LEVEL II

UNITED STATES DEPARTMENT OF COMMERCE • Maurice H. Stans, *Secretary*
NATIONAL BUREAU OF STANDARDS • A. V. Astin, *Director*

Precision Measurement and Calibration

Volume I.

~~Selected NBS Papers~~

Statistical Concepts and Procedures

Harry H. Ku, Editor

A compilation of previously published papers by the staff of the National Bureau of Standards, including selected abstracts by NBS and non-NBS authors. Issued in several volumes, see page IV.

NBS Special Publication 300 — Volume 1

Issued February 1969

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. - Price $10.25

Stock No. 003-003-00072-8

Abstract

This volume is one of an extended series which brings together the previously published papers, monographs, abstracts, and bibliographies by NBS authors dealing with the precision measurement of specific physical quantities and the calibration of the related metrology equipment. The contents have been selected as being useful to the standards laboratories of the United States in tracing to NBS standards the accuracies of measurement needed for research work, factory production, or field evaluation.

Volume 1 deals with methodology in the generation, analysis, and interpretation of precision measurement data. It contains 40 reprints assembled in 6 sections: 1) The Measurement Process 2) Design of Experiments in Calibration 3) Interlaboratory Tests 4) Functional Relationships 5) Statistical Treatment of Measurement Data 6) Miscellaneous. Each section is introduced by an interpretive foreword, and the whole is supplemented by abstracts and selected references.

Key Words: Accuracy; analysis of measurement data; design of experiments; functional relationships; interlaboratory tests; measurement process; precision; statistical concepts in measurements; systematic error.

Library of Congress Catalog Card Number: 68-60042

Foreword

In the 1950's the tremendous increase in industrial activity, particularly in the missile and satellite fields, led to an unprecedented demand for precision measurement, which, in turn, brought about the establishment of hundreds of new standards laboratories. To aid these laboratories in transmitting the accuracies of the national standards to the shops of industry, NBS in 1959 gathered together and reprinted a number of technical papers by members of its staff describing methods of precision measurement and the design and calibration of standards and instruments. These reprints, representing papers written over a period of several decades, were published as NBS Handbook 77, Precision Measurement and Calibration, in three volumes: Electricity and Electronics; Heat and Mechanics; Optics, Metrology, and Radiation.

Some of the papers in Handbook 77 are still useful, but new theoretical knowledge, improved materials, and increasingly complex experimental techniques have so advanced the art and science of measurement that a new compilation has become necessary. The present volume is part of a new reprint collection, designated NBS Special Publication 300, which has been planned to fill this need. Besides previously published papers by the NBS staff, the collection includes selected abstracts by both NBS and non-NBS authors. It is hoped that SP 300 will serve both as a textbook and as a reference source for the many scientists and engineers who fill responsible positions in standards laboratories.

A. V. ASTIN, *Director*

Preface

The general plan for this compilation has been reviewed by the Information Committee of the National Conference of Standards Laboratories. The plan calls for Special Publication 300 to be published in 12 volumes having the following titles and editors:

Statistical Concepts and Procedures, H. H. Ku
Frequency and Time, A. H. Morgan
Electricity—Low Frequency, F. L. Hermach and R. F. Dziuba
Electricity—Radio Frequency, A. J. Estin
Heat, D. C. Ginnings
Temperature, J. F. Swindells
Mechanics, R. L. Bloss
Dimensional Metrology—Length and Angle, H. K. Hammond, III
Radiometry and Photometry, H. K. Hammond, III
Colorimetry and Image Optics, H. K. Hammond, III
Spectrochemical Analysis, B. F. Scribner
Ionizing Radiation, E. H. Eisenhower

This division of subject matter has been chosen to assure knowledgeable selection of context rather than to attain uniform size. It is believed, however, that the larger volumes, of approximately 600 pages, will still be small enough for convenient handling in the laboratory.

The compilation consists primarily of original papers by NBS authors which have been reprinted by photoreproduction, with occasional updating of graphs or numerical data when this has appeared desirable. In addition, some important publications by non-NBS authors that are too long to be included, are represented by abstracts or references; the abstracts are signed by the individuals who wrote them, unless written by the author.

Each volume has a subject index and author index, and within each volume, contents are grouped by subtopics to facilitate browsing. Many entries follow the recent Bureau practice of assigning several key words or phrases to each document; these may be collated with titles in the index. Pagination is continuous within the volume, the page numbers in the original publications also being retained and combined with the volume page numbers, for example 100-10. The index notation 1-133 refers to volume 1, page 133 of this volume. A convenient list of SI (Système International) physical units and a conversion table are to be found inside the back cover.

The publications listed herein for which a price is indicated are available from the Superintendent of Documents, U. S. Government Printing Office, Washington, D. C. 20402 (foreign postage, one-fourth additional). Many documents in the various NBS nonperiodical series are also available from the NBS Clearinghouse for Federal Scientific and Technical Information, Springfield, Va. 22151. Reprints from the NBS Journal of Research or from non-NBS journals may sometimes be obtained directly from an author.

Suggestions as to the selection of papers which should be included in future editions will be welcome. Current developments in measurement technology at NBS are covered in annual seminars held at either the Gaithersburg (Maryland) or the Boulder (Colorado) laboratories. These developments are summarized, along with a running list of publications by NBS authors, in the monthly NBS Technical News Bulletin.

H. L. MASON,
Office of Measurement Services
NBS Institute for Basic Standards.

Editor's Note

This volume deals with methodology in the generation, analysis, and interpretation of precision measurement data. It is a collection of papers that have been found useful to the measurement fraternity, as represented by participants in the annual NBS seminars on Precision and Accuracy in Measurement and Calibration. The main criterion used in selection was ease of communication; that is, whether the author's message gets across to the general reader, so that he can develop the idea for gainful application in his own specialized area.

The volume contains reprints of 40 papers on statistical concepts and procedures classified in six sections. Four works too long to be included here are represented by titles and abstracts in Section 7. The interpretive foreword appearing at the beginning of each of the first six sections comments on the individual papers and thus characterizes the particular section. The index has been prepared to facilitate browsing. Paper 6.8 provides a list of selected references, annotated for the reader's convenience. Some of these are referred to in the various forewords.

I wish to acknowledge my indebtedness to Churchill Eisenhart and to members of the Statistical Engineering Laboratory for their suggestions in selection of papers, and for their help in the preparation of this volume.

Thanks are also due to publishers of non-NBS papers for permission to reprint in this volume papers by D. B. De Lury, William H. Kruskal, R. B. Murphy, Milton Terry, and E. Bright Wilson, Jr.

HARRY H. KU, *Editor*

Contents

3. Interlaboratory Tests

4. Functional Relationships

5. Statistical Treatment of Measurement Data

1. The Measurement Process, Precision, Systematic Error, and Accuracy

Foreword

Statistical control on the quality of manufactured items formally began with Walter Shewhart some forty years ago, but statistical control on the quality of precise measured values in a calibration laboratory did not become a reality until just recently. The first published example of this realization appears to be that given by Pontius and Cameron in their Monograph (1.1) on mass measurement.

A prime mover in the transfer of this basic concept from production processes to measurement processes has been Churchill Eisenhart, who has spent much of his time the last two decades advocating this discipline both within and without the Bureau. A definitive treatise based on his study appears as the second paper, Realistic Evaluation (1.2).

The "postulate of measurement," which Eisenhart used in his paper and which he attributed to N. Ernest Dorsey, originated from Dorsey's treatise, The Velocity of Light. Excerpts from this work of Dorsey's, selected and arranged by Eisenhart, are reprinted here under the title, On Absolute Measurement (1.3).

In Systematic Errors in Physical Constants (1.4), Youden extended Dorsey's observations on the effects of changing environmental conditions, and introduced the use of weighing designs into physical experimentation. These designs, labeled as Youden's ruggedness test designs in his papers in section 3, are constructed for the efficient and systematic searching out of systematic errors.

Youden's other paper (1.5) emphasized the use of statistical design to get an indirect estimate of the error in comparing an instrument with a reference standard. He pointed out that users of calibrated items often have an optimistic notion of the quality of the measurements they make, and suggested that some investigation should be made in order to ascertain whether some of the demands made for better standards are justified.

The presentation of final results, and the uncertainties associated with the realizations of the measurement method by which these results are obtained, has always been a source of difficulty. The recommendations given in the Expression of the Uncertainties of Final Results (1.6) and the tabular guide to commonly used terms and expressions (1.7) are included to serve as references to experimenters who are faced with this problem.

UNITED STATES DEPARTMENT OF COMMERCE
Alexander B. Trowbridge, *Secretary*
NATIONAL BUREAU OF STANDARDS • A. V. Astin, *Director*

Realistic Uncertainties and the Mass Measurement Process

An Illustrated Review

P. E. Pontius and J. M. Cameron

Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234

The illustrations were regrouped to improve the format. July 1968.

National Bureau of Standards Monograph 103

Issued August 15, 1967

Realistic Uncertainties and the Mass Measurement Process

An Illustrated Review

Paul E. Pontius and Joseph M. Cameron

This paper gives a review of the concepts and operations involved in measuring the mass of an object. The importance of viewing measurement as a production process is emphasized and methods of evaluating process parameters are presented. The use of one of the laboratory's standards as an additional unknown in routine calibration provides an accuracy check and, as time goes on, the basis for precision and accuracy statements.

Key Words: Measurement, measurement process, uncertainty, mass measurement, precision, accuracy, statistical control.

Introduction

This paper is a condensed version of a lecture on "Error of Measurement" presented by Paul E. Pontius and Joseph M. Cameron at the Seminar on Mass Measurement, held at the National Bureau of Standards, Washington, D. C., November 30, December 1 and 2, 1964, and is essentially as presented by Paul E. Pontius at the 20th Annual ISA Conference held at Los Angeles, California, October 4–7, 1965.

It is a review of the mass measurement process from the initial basic concept to the statement of a measured mass value, examining in more or less detail certain important elements which are apt to be misunderstood, or perhaps misused. The importance of viewing measurement as a production process is emphasized and methods of evaluating process parameters are presented. The use of one of the laboratory's standards as an additional unknown in routine calibration provides an accuracy check and, as time goes on, the basis for precision and accuracy statements.

Mass Measurement Requirements

One role of the Bureau is to provide an extension of the mass measurement unit into the facilities of those who must use mass values to do other useful work. . . .

Figure 1

These large weights, for example, are for use by another part of the Bureau to calibrate force measuring devices.

★★★

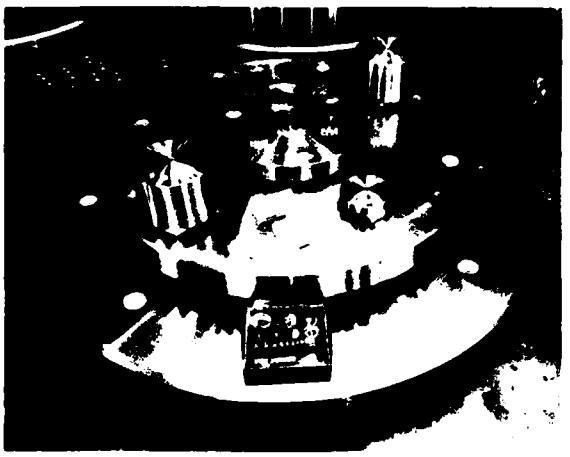


Figure 2

The calibration service provides values for single, selected groups, and ordered sets of standards, the values being with reference to the national standard of mass. These values, together with a value for their uncertainty, allow each user to determine, in combination with his measurement process, the uncertainty of his measurements.

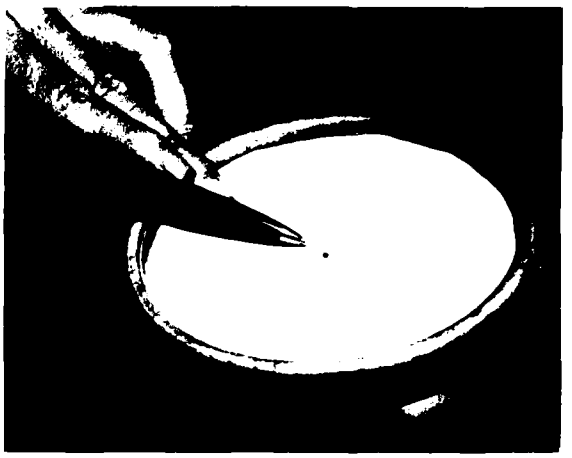


Figure 3

The three photographs above started with a group of standards whose cumulative total mass was in excess of one million pounds, and ends with a micropound standard, a range in excess of ten to the twelfth power (10^{12}).

★★★

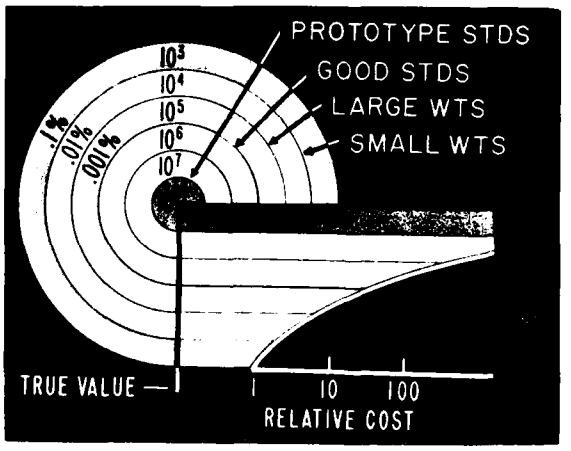


Figure 4

The accuracy requirements for a measurement are set partly by experience, partly by discussions with others, and partly by analysis. For a particular purpose, the accuracy requirement must be established with care, as it provides a point of departure for the entire measurement process. Frequently we tend to lose perspective in regard to what we are measuring, or what the measurements mean, particularly if we concentrate on routine procedures or are remote to the actual measurement.

The aiming point for our measurement is to establish the mass, or true value, of a particular object for it is, in concept at least, unique and invariant. If, for example, accuracy within .01 percent is sufficient for our purpose, the target center is the area within the next to the last circle. Our measurements may group on either side of dead center, or may be randomly scattered across the center of the target, but as long as the spread is essentially within the target circle, the process is satisfactory for its intended use. Troubles arise when realistic requirements are divided by large arbitrary constants as specifications pass through various groups of people in a complex organization. Measurements accurate to better than .01 percent require attention to many details under more or less ideal conditions, and may not be obtainable under adverse conditions, consequently the entire measurement effort may be lost if the end use involves measurement processes of questionable precision. In the case of calibration, for example, in order to utilize the accuracy inherent in a good calibration, the user must work just as hard in his measurement process as the calibration facility did to determine the value of the standard originally.

★ ★ ★

The importance of incorporating the properties of the measurement process in setting up requirements or specifications is illustrated by the problem of adjustment tolerances for different classes of weights.

	TYPICAL PROCESS PARAMETERS			CLASS ADJ. TOL.	
NOMINAL VALUE	UNCERTAINTY (SYS. ERROR) OF STD. VALUE	S.D. OF SINGLE MEAS.	SINGLE MEAS. PROCESS UNCERTAINTY*	CLASS M (mg)	CLASS S (mg)
10 g	.0087mg	.0074mg	.031mg		.074
5 g	.0050	.004	.017		.054
1 g	.0047	.004	.017		.054
500 mg	.0024	.0007	.005		.025
100 mg	.0009	.0007	.003	.010	.025
10 mg	.0008	.0007	.003	.010	.014

*3 times one standard deviation of the measurement process plus bound to possible systematic errors.

Figure 5

The Class M and Class S adjustment tolerance limits for selected weights are shown in the two right hand columns. The uncertainty associated with the stated value for standards of the same nominal value is shown in the 2d column and the precision for a single measurement is shown in the 3d column. If one tries to establish the compliance with Class M adjustment tolerances by a single weighing against a known standard, the uncertainty of the process would be as shown in the 4th column. This uncertainty, compared with the quantity we are trying to detect, is such that in the first 4 cases the measurement uncertainty is a large fraction of the tolerance so that only those items well inside of tolerance have a good chance of being passed. A measurement procedure more sophisticated than a single comparison with a known standard may be desirable.

★ ★ ★

	TYPICAL PROCESS PARAMETERS			CLASS ADJ. TOL.	
NOMINAL VALUE	UNCERTAINTY OF CLASS M (WITHIN TOL.)	S.D. OF SINGLE MEAS.	SINGLE MEAS. PROCESS UNCERTAINTY	CLASS S (mg)	CLASS S-1 (mg)
10 g	.050	.0074	.072		.18
5 g	.034	.004	.048		.18
1 g	.034	.004	.048		.10
500 mg	.010	.0007	.012	.025	.08
100 mg	.010	.0007	.012	.025	.05
10 mg	.010	.0007	.012		.03

Figure 6

We would be in greater difficulties if we were to try to establish compliance with Class S adjustment tolerances in the same manner with reference to Class M standards, which are known only to be within the Class M tolerance limits. In 4 of the 6 examples, the process uncertainty is of the same order of magnitude as the quantity we are trying to check. These examples illustrate the necessity for a careful evaluation before venturing a commitment on the performance of a particular measurement process.

★ ★ ★

The Unit of Mass

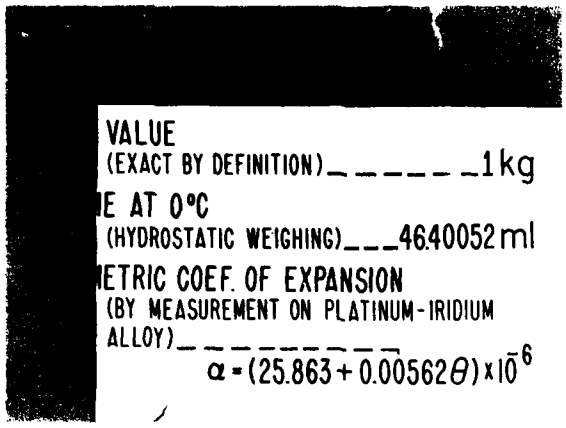


Figure 7

By practically universal agreement, the mass of the International Prototype Kilogram is the basic unit for mass measurement. It is a particular object, defined to have an exact invariant mass of one kilogram, that is to say, the true value is one kilogram. The volume and the coefficient of volumetric expansion are necessary to determine the best estimate of the true value of other objects compared with this standard.

★ ★ ★

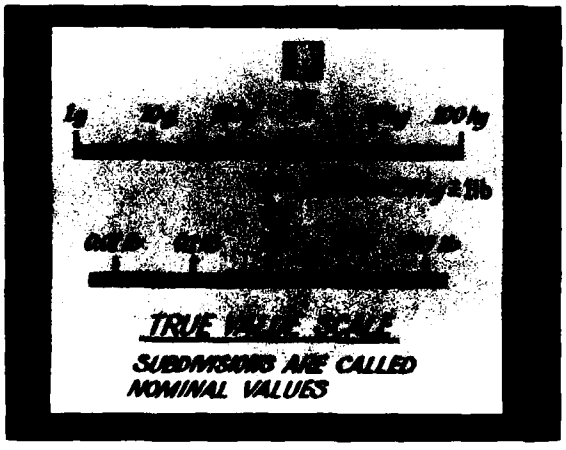


Figure 8

With the unit defined, we can logically construct a true value scale which has the property that some point on the scale will correspond to the mass of any chosen object. We call the major subdivisions of this scale nominal values. Other customary units, such as the pound, are not ambiguous if they have an exact definition relative to the basic unit. An intermediate point on the scale can be described either relative to the whole scale, as for example, 9.995 grams, or relative to the closest nominal value, in which case the point would be described as 10 grams minus 5 milligrams. The minus 5 milligrams may be called a correction or error, depending on one's viewpoint. The use of a nominal value and a correction is often convenient in computations, however, the word "correction", or "error", overly emphasizes the importance of the nominal value. Interpretation of tolerance limits on the value of the standard as the error automatically disregards the primary benefits of a good calibration. Only an ideal measurement method or process can produce true values of multiples and subdivisions of the basic unit which will exactly coincide with nominal values on the true value scale. It should be emphasized that, from a measurement standpoint, adjustment to nearly coincide with a nominal value is necessary only to assure an "on scale" condition when intercomparing equal nominal summations.

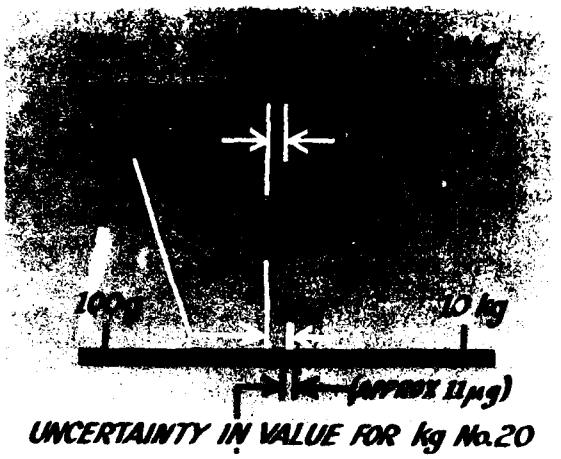
★ ★ ★

	TYPICAL PROCESS PARAMETERS			CLASS ADJ. TOL	
NOMINAL VALUE	UNCERTAINTY (SYS. ERROR) OF STD. VALUE	S.D. OF SINGLE MEAS.	SINGLE MEAS. PROCESS UNCERTAINTY *	CLASS S (mg)	CLASS S-1 (mg)
10 g	.0087 mg	.0074 mg	.031 mg		.18
5 g	.0050	.004	.017	.054	.18
1 g	.0047	.004	.017	.054	.10
500 mg	.0024	.0007	.005	.025	.08
100 mg	.0009	.0007	.003	.025	.05
10 mg	.0008	.0007	.003	.014	.03

*3 S.D. + SYS. ERROR

Figure 9

In our previous example, we elected to interpret the adjustment tolerance limits associated with our Class M set as the uncertainty of the value. While this may be appropriate with respect to the nominal value, such an interpretation raised serious doubts as to our ability to test the Class S weight set. If we had used the actual value and its uncertainty as a basis for our tests, the doubt essentially disappears. With minor modification at the 10 g level, the uncertainty of the values established for the Class S weights by our single measurement is clearly suitable for the task at hand. It must be emphasized that our apparent increase in measurement capability did not require any change in our process hardware. It has been achieved, for the most part, by a change in philosophy.

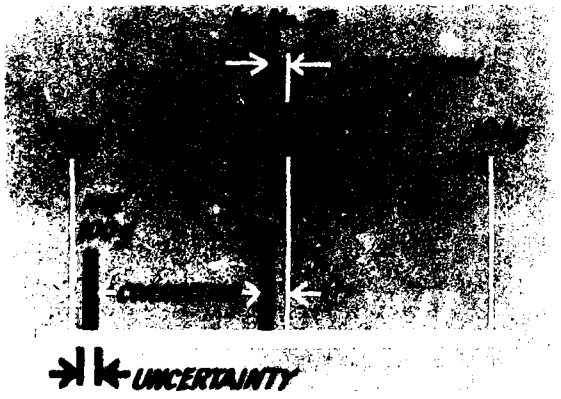


UNCERTAINTY IN VALUE FOR kg No.20

Figure 10

Our access to the true value scale as established by the international standard is through prototype kilogram number 20. The estimated true value of number 20 is 1 kilogram minus 19 micrograms, based on several measurements. We can construct an accessible true value scale by setting off from the value of kg 20 an amount equal to the correction. Practically, the stated value is assumed to be exact, the uncertainty of the value introducing only a slight systematic error in our reconstructed scale.

★ ★ ★



TRUE VALUE OR IDEAL SCALE

Figure 11

By comparing other objects with kilogram 20, either singly or in combination, we can assign values relative to our accessible scale. A sufficient number of well calibrated standards which can be intercompared, and which may occasionally be compared with our prototype standard, serve to maintain our scale with perhaps a greater precision than was available in the starting measurements. All mass values on NBS Reports of Calibration are with reference to a minimum number of selected mass standards. For example, practically all sets of metric weights are calibrated with reference to a pair of 1 kg or a pair of 200 g or a pair of 100 g weights. The national reference standards group does *not* include weights of all denominations.

★ ★ ★

Measurement Method

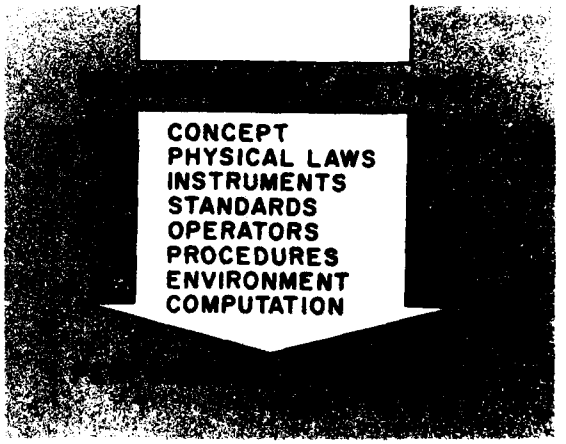


Figure 12

A practical measurement method is easy to visualize in the form of a broad outline of the elements of the method such as, the concept of the quantity to be measured, pertinent physical laws, various instruments, standards, the operators, procedures to be used, the environment in which the measurements are to be made, the computations which are to be made, and a means of establishing some parameters of performance. As we briefly review some of these elements, we will find that every mass measurement facility has many things in common.

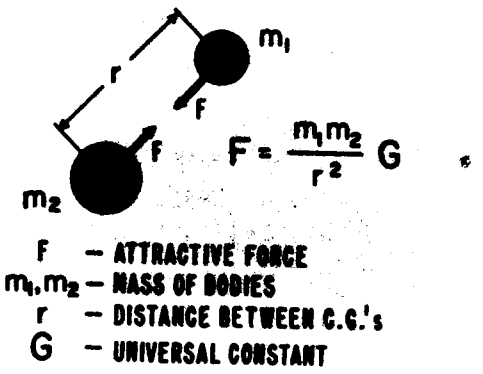


Figure 13

Mass is an inertial property of an object, which, within the framework in which our measurements apply, is considered to be proportional to the amount of material. Mass is generally thought of as being measured through some application of Newton's law of gravitational attraction, however, it is perhaps more precise to say that measurements are made by comparing the forces attracting suspended bodies toward the earth—that is the net vertical forces including the effects of G, air buoyancy, rotation of the earth, etc.

★★★

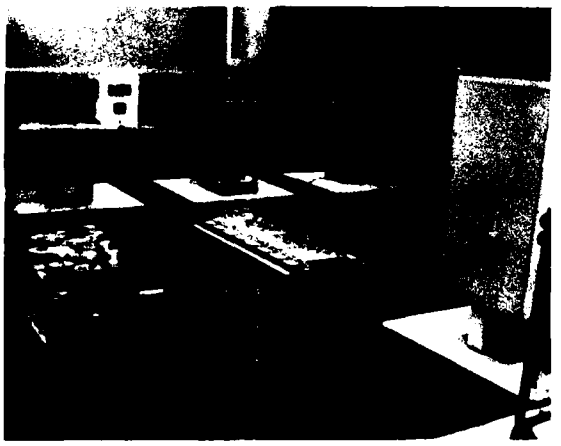


Figure 14

The environment in which the measurements are made does not vary substantially between calibration facilities. Weighing rooms are almost universally clean, with restricted access, and relatively free of vibration. With the possible exception of freedom from vibration, these desirable features are easily obtained.

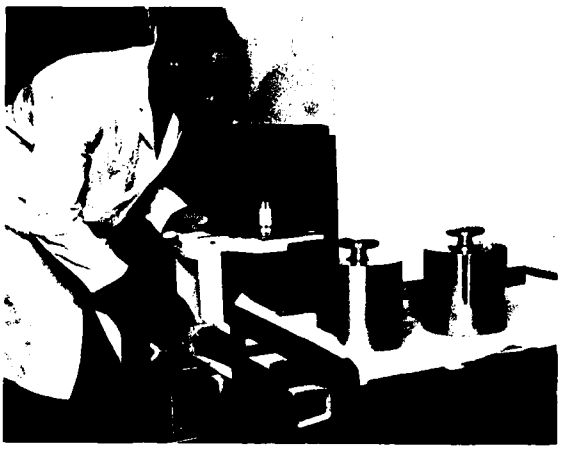


Figure 15

People operate the equipment, following prescribed procedures. Operator skill increases with practice, and in time, operators in a given group approach a uniform level of skill.

★★★

Each comparison, or weighing, consists of a sequence of operations, more or less formalized. Detailed procedures and weighing designs, ranging from simple to complex, are available for a wide variety of requirements. Modern computation equipment ranging from desk calculator to electronic computer are now widely available so that laborious long hand computations are no longer necessary.

While perhaps not generally considered so, analysis is a part of the measurement method. Whether done by machine . . .

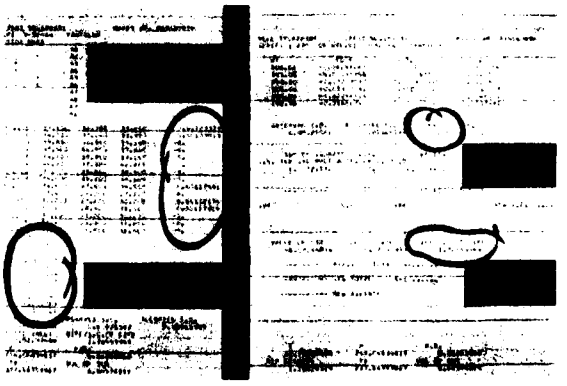


Figure 16

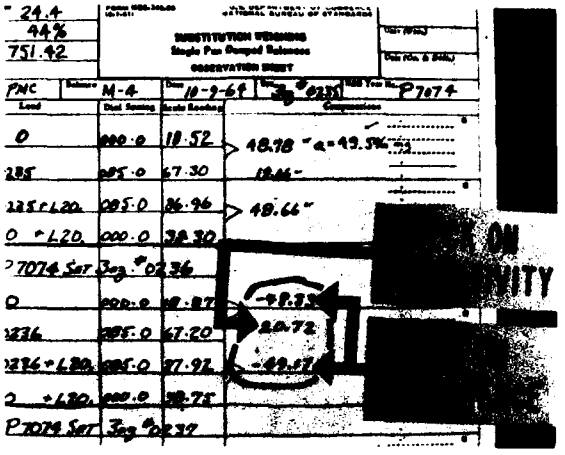


Figure 17

. . . or by hand, the analysis verifies that such parameters continue to be applicable.

★ ★ ★

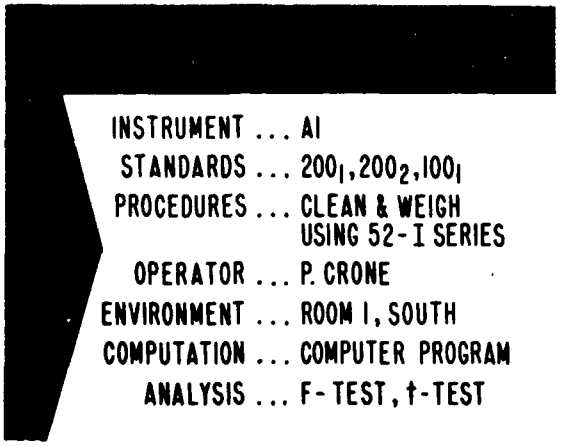


Figure 18

A particular measurement method is like a specification for a particular measurement. The specific instrument, the standards to be used, the specific operations to be performed and the planned sequence in which they are to be carried out, the operator, the location, and the method of computation and analysis, collectively define a particular measurement method. Until the measurement has actually been made and analyzed, the performance is only "on paper" and therefore ideal.

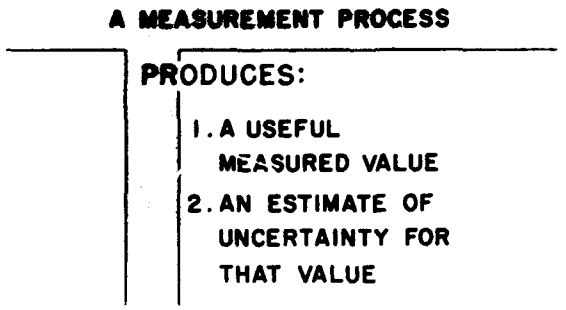


Figure 19

A measurement process involves the actual physical operation of the specified equipment following the procedures as closely as possible. It is subject to the many variations that can and do occur during the operation. The end result is an estimated best value, which, in order to be useful, must be accompanied by the uncertainty with reference to known performance parameters.

Changes in any one or in a group of elements of the method constitutes, in effect, a different particular method and a different process which will in turn produce a different result and a different uncertainty. Small changes can make the difference between a useful value or a wasted effort.

★ ★ ★

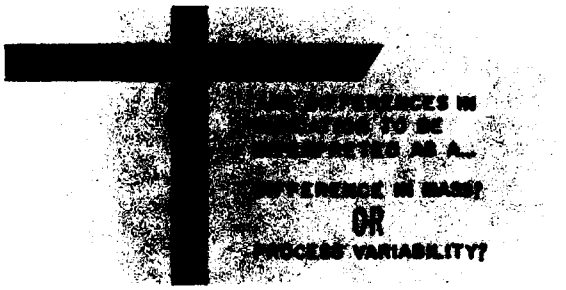


Figure 20

Because we must establish the mass of the object in question by measuring the mass difference between it and some known standard, the comparator is a vital element in the process. The inherent characteristic of the comparator is precision—not accuracy. The fundamental question is whether the indicated difference is really a mass difference, or an indication of some other variability. While we may be able to identify large sources of variability, in the limit, we cannot differentiate between instrument precision, variability from extraneous sources, or variability of the standard.

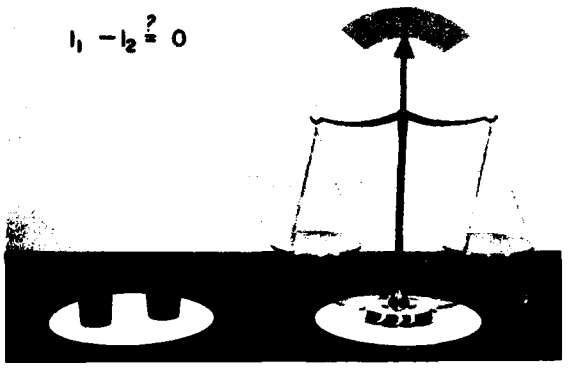


Figure 21

We start by determining the indicated difference between two objects that are nearly alike.

★ ★ ★

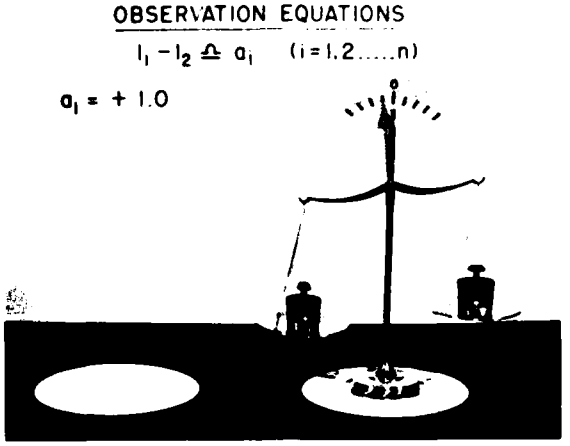


Figure 22

From our first comparison, it appears that the round knob weight on the left is clearly heavier than the flat knob weight by one scale division. If we stop here, we would simply state the value of one object in terms of another, however, we have no way of knowing the uncertainty to associate with this value.

(The symbol $\triangleq$ signifies that the relationship is not a strict equality because of the random errors of measurement that are present on the right side.)

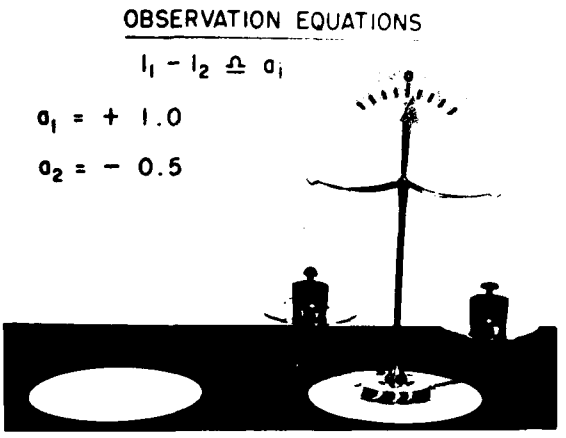


Figure 23

If we repeat the comparison at some other time, we are quite likely to obtain a different result. This raises a serious question—which of the two results is correct?

★ ★ ★

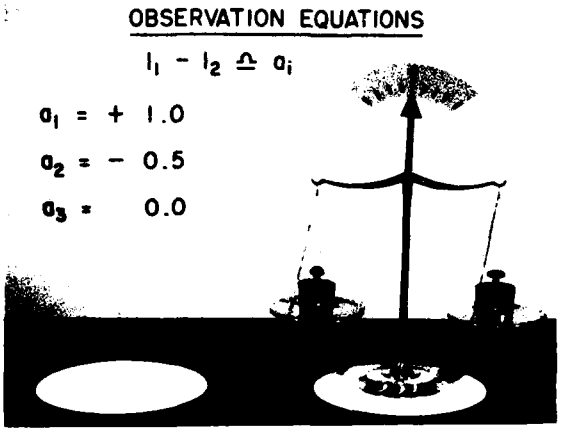


Figure 24

We repeat the comparison again . . .

★ ★ ★

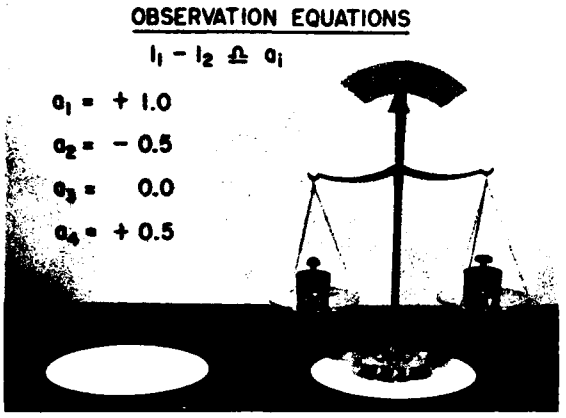


Figure 25

. . . and again. Now there are four different values, none of which alone can be considered the best measure of the difference, but considered as a group they can tell us something about the instrument. Continuing to record the indicated difference between two similar objects, and preferably making the comparisons in the environment in which the instrument is to be used, a plot is made against time of the differences which may look like this.

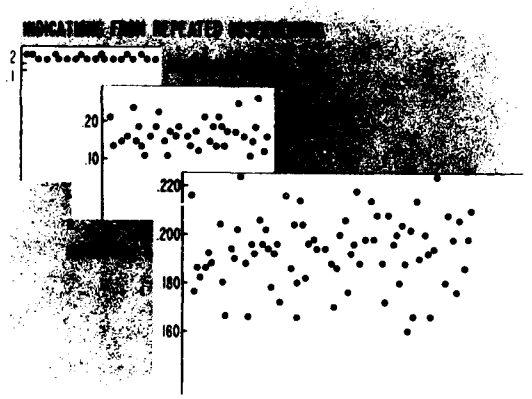


Figure 26

The first plot indicates a severe rounding off, which may be from several causes. Such a response clearly lacks the appearance of randomness. The second plot at least appears to be random. The third plot, while perhaps appearing to be random, obviously lacks the precision of the second plot. The range of the differences as plotted gives us an idea of the smallest mass difference that can be detected with assurance, and is obviously related to the requirements our measurements must meet. Repeated independent measurements of the same mass difference are essential to the evaluation of the instrument.

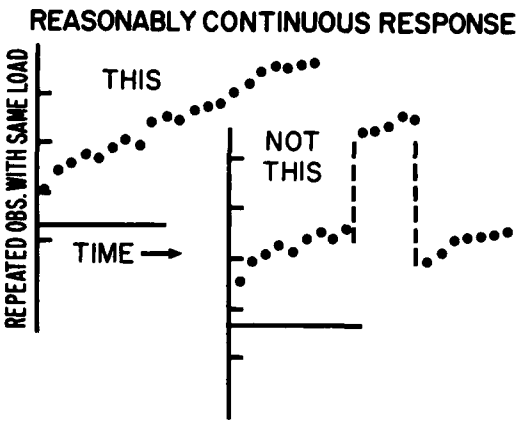


Figure 27

The operator, or manufacturer, must search for cause and effect until repeated indications for the same load, or differences are reasonably consistent. Effects which are periodic in nature, but with a period significantly longer than the period of the instrument, can be minimized in the design of the weighing method.

★ ★ ★

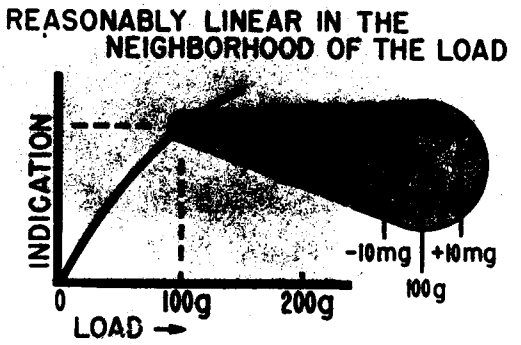


Figure 28

One additional requirement, generally beyond the control of the operator, is that of linearity. An instrument, used as a comparator rather than a direct reading device, requires linearity only in the neighborhood of the actual load.

OBSERVED DIFFERENCES TO MASS DIFFERENCES

1. SUBSTITUTION
2. TRANSPOSITION
3. "DIRECT READING"

Figure 29

The problem of establishing the correspondence between observed differences and mass differences is a part of the weighing method. The first two methods, substitution and transposition, are comparative methods. That is to say, the method requires observations relative to a suitable standard along with the unknown. With these methods, the measurement equipment need be continuous only over the time interval required for making a group of observations and linear only over the range of the difference between the standard and the unknown. Most direct reading equipment is in a sense a substitute standard, that is, at some point in time it is calibrated with reference to a standard, and from that point until recalibration, it is generally assumed to have a long term constancy approaching that of the standard. Most mass measurement equipment can be used either way. The smallest uncertainties invariably will be associated with the comparative mode of operation.

★ ★ ★

SUBSTITUTION METHOD

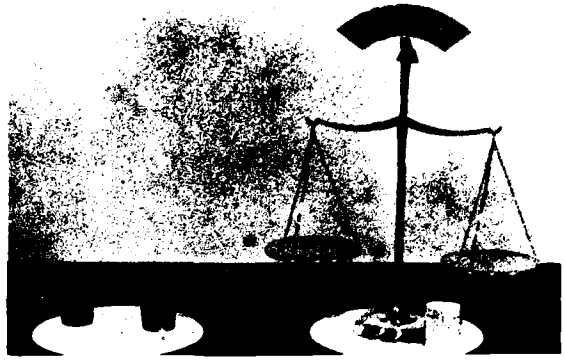


Figure 30

To illustrate the principle, the double substitution method is performed as follows: We start with a simulated equal arm balance, a tare weight—the white cylinder near the base of the balance, a sensitivity weight of known value immediately in front of the dark weight near the center, and two nearly equal brass weights, one with a flat knob in the center and one with a round knob on the left. The scale indication is in arbitrary numbers and the tare weight is necessary to establish an "on scale" condition.

★ ★ ★

(I) $A \rightarrow O_1$

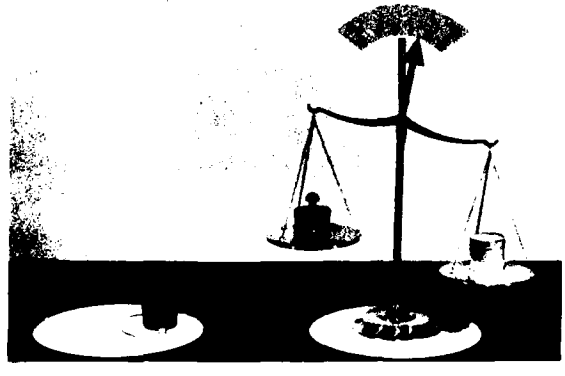


Figure 31

The first observation is that produced with the round knob weight on the pan.

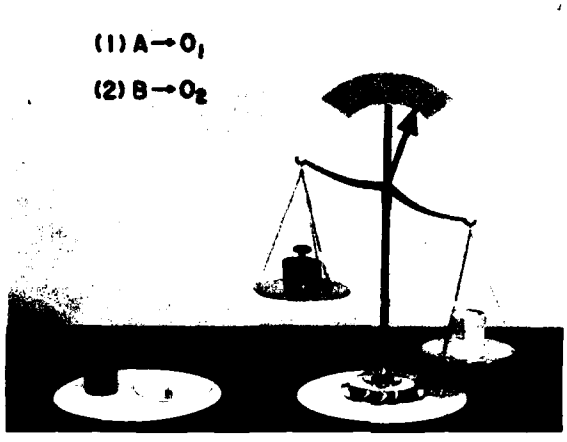


Figure 32

The second observation is that produced with the flat knob weight, which might be a standard, replacing, or substituted for, the round knob weight.

★ ★ ★

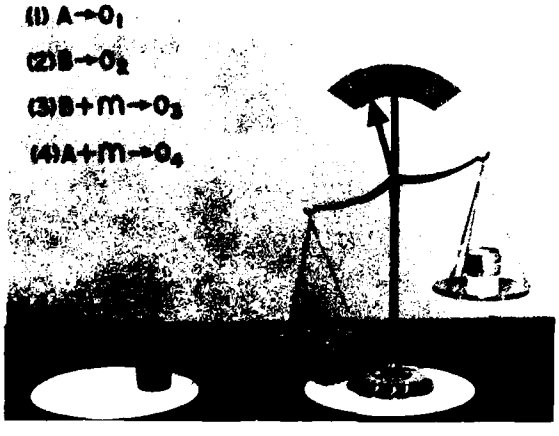


Figure 33

The third observation is that produced by repeating the previous step and adding the sensitivity weight to the pan load.

★ ★ ★

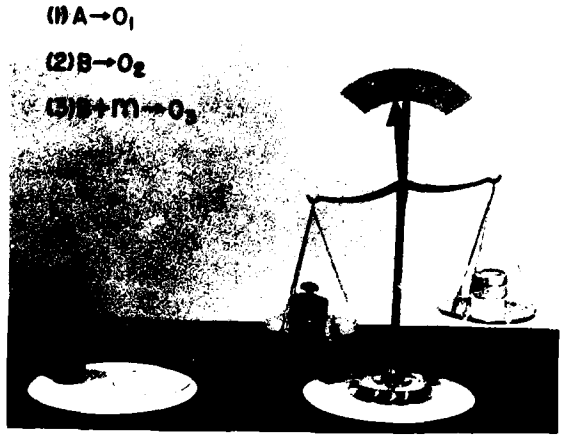


Figure 34

The fourth observation is a repetition of the first step including the sensitivity weight.

★ ★ ★

$$A-B \triangleq K\left\{\frac{(1)-(2)+(4)-(3)}{2}\right\}$$

$$\triangleq K\left(\frac{O_1-O_2+O_4-O_3}{2}\right)$$

$$B+m-B \triangleq K\{(3)-(2)\}$$

$$m \triangleq K(O_3-O_2)$$

$$A-B \triangleq \frac{m}{(O_3-O_2)}\left(\frac{O_1-O_2+O_4-O_3}{2}\right)$$

Figure 35

Using the requirement for continuity, a relation can be established for A minus B from the average of the two sets of differences as shown. Using the linearity requirement, the constant of proportionality K, or the mass value of the indicating scale division can be determined from the second and third observation. Finally, the difference A minus B is expressed as a function of the observations, in ratio form and the value of the sensitivity weight.

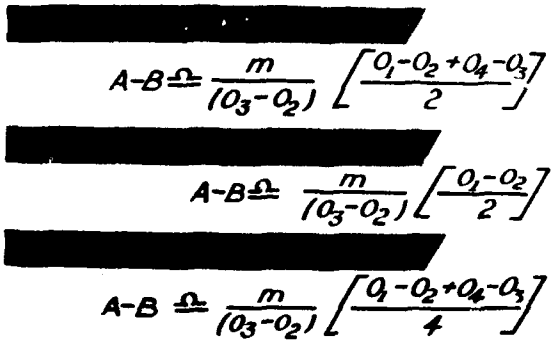


Figure 36

All usual methods result in very similar relations expressing the difference between two objects being compared. In all cases, A minus B is expressed as a ratio between sets of observations multiplied by the value of the sensitivity weight. Obviously requirements for knowledge of the value of m are minimized when the size of the ratio involving the observation is small. The constant of proportionality, K, is really the ratio in front of the bracket terms which we call the value of the division. The strange equal sign is used to indicate that the relations shown are observational equations and not mathematical identities.

★★★

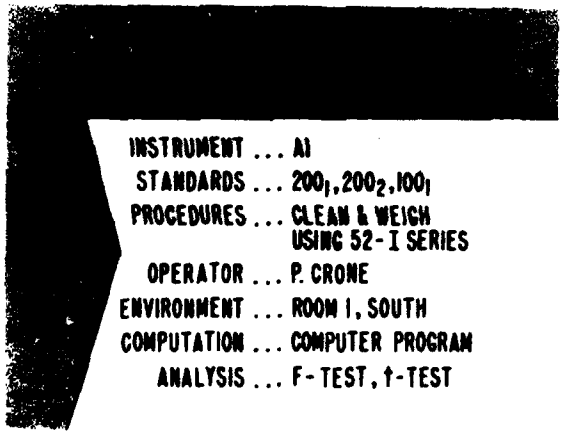


Figure 37

With the measurement method agreed upon, let us now discuss its performance—we put it into production and see how it works out as a measurement process.

Measurement as a Process

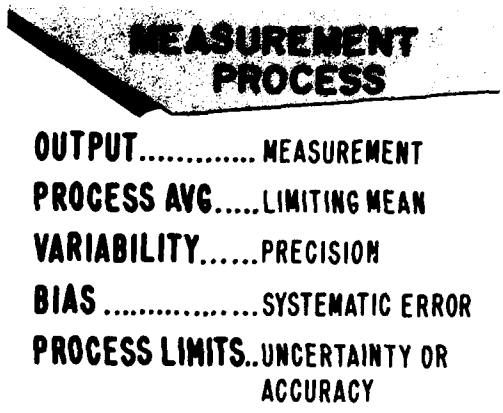


Figure 38

A measurement process is essentially a production process, the "product" being numbers, that is, the measurements. A characteristic of a measurement process is that repeated measurements of the same thing result in a series of non-identical numbers. To specify a measurement process involves ascertaining the limiting mean of the process: its variability due to random imperfections in the behavior of the system, that is, its precision: possible extent of systematic errors from known sources, or bias; and overall limits to the uncertainty of independent measurements.

★★★

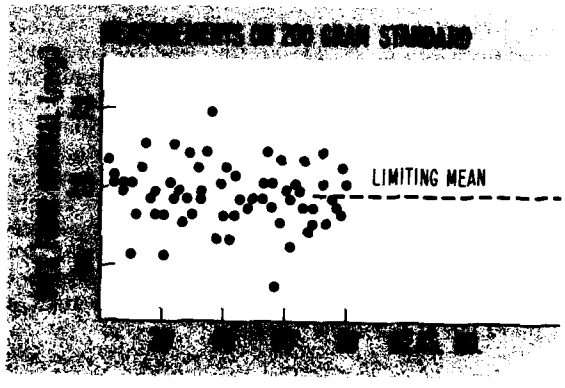


Figure 39

The chart shows measurements on a 200 g weight, plotted in the order in which they were taken. Despite the presence of one or two stragglers, the measurements tend to cluster around the central line—the process average or limiting mean. Our confidence that the process has settled down to a single limiting mean is strengthened as the length of the record is increased. We may have satisfied ourselves regarding the mean but what about the next measurement?

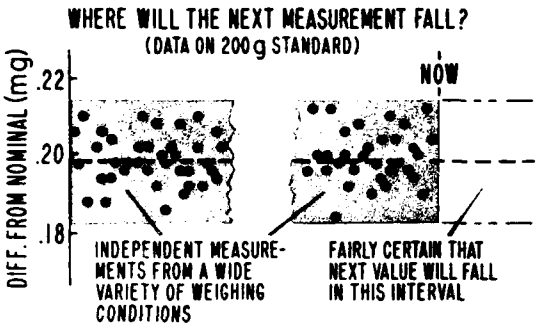


Figure 40

It seems clear that we cannot give an exact answer but will have to content ourselves with a statement that allows for the scatter of the results. Our goal is to make a statement with respect to a new measurement that is independent of all those that have gone before. As indicated in the chart, if we had a sufficiently long record of measurements we could set limits within which we were fairly certain that the next measurement would lie. Such a statement should be based on a collection of independent determinations, each one similar in character to the new observation, that is to say, so that each observation of the collection and also the new observation can be considered as random drawings from the same probability distribution. These conditions will be satisfied if the collection of points is independent, that is free of patterns, trends and so forth: and provided it is from a sufficiently broad set of environmental and operating conditions to allow all the random effects to which the process is subject, to have a chance to exert their influence on the variability. Suitable collections of data can be obtained by incorporating an appropriate measurement into daily routine weighing procedures, for example, a daily measurement of the difference between two laboratory weights, or in the regular calibration of the same weight.

★★★

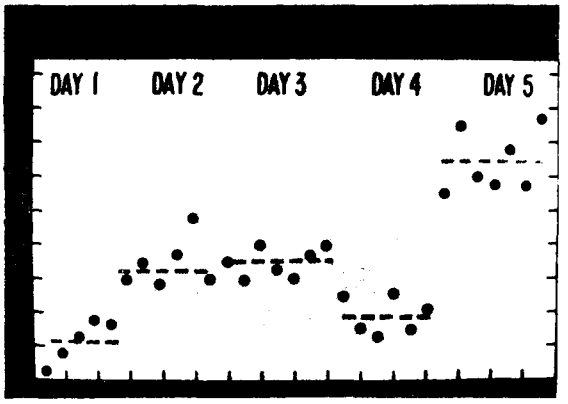


Figure 41

If the measurements tend to cluster when taken close together in time, like the results shown on the chart, some systematic effect is present and certainly the results are not independent. This may be due to some as yet undetermined cause, and the group means may have the appearance of randomness of the previous chart.

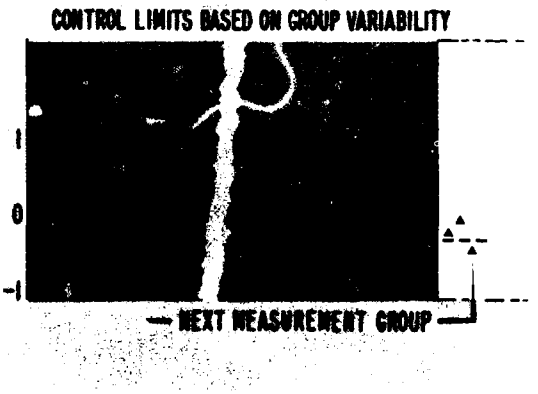


Figure 42

The group means may tend to a limit and the process may have all the properties of a good measurement system, once the allowance is made for the grouping. It is important that grouping be properly handled in determining the precision of the process. By modifying the process or changing the schedule of measurements to give the effect of independent measurements, we can arrive at a situation like the values on the 200 g standard. The shaded band is meant to suggest a limit, not an artistic slide.

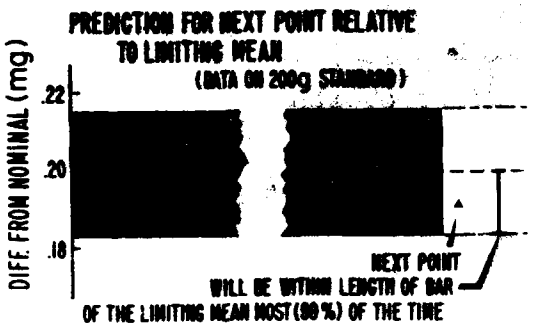


Figure 43

From a study of a sequence of such independent measurements, we can use control chart techniques to set up limits within which the next value should lie. In the case where we have an extremely long sequence, a bar, as illustrated in the chart, can be marked off on either side of the mean so that some suitable fraction, say 99 percent, of the observations are within the interval represented by its length.

★ ★ ★

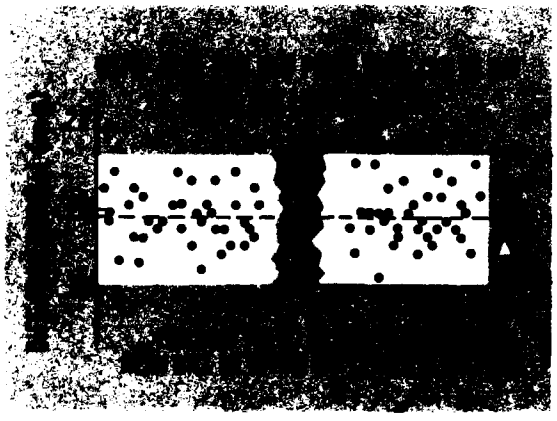


Figure 44

We can reverse the process and say that the probability is 99 percent, that the true value, or limiting mean, will not be more than the width of the bar from any observation chosen at random. This will be true of the next observation as well, provided it is an independent measurement from the same process. The probability statement attaches to the sequence of such statements. For each individual new observation the statement is either true or false but in the long run 99 percent of such statements will be true.

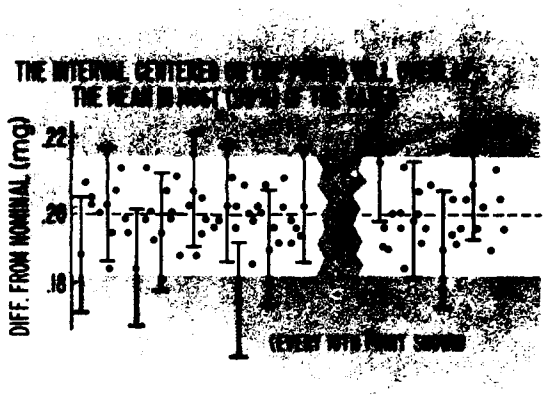


Figure 45

Assuming that the limits on the chart are based on large numbers of observations, we would find that very nearly the intended percentage of all such bars, centered on the observed values, would in fact overlap the mean. Only in those cases, such as the points in the area outside of the control limits, will the bar fail to overlap the mean. This is expected in only 1 percent of the cases. More frequent occurrence is a clear indication of either loss of control or that the limits were not properly set. Once we are satisfied that the process has a limiting mean value and is stable enough to permit prediction we turn our attention to evaluating its precision.

★ ★ ★

Process Precision

Let us now take a look at the situation in weighing to see what is involved in the study of the precision of the process.

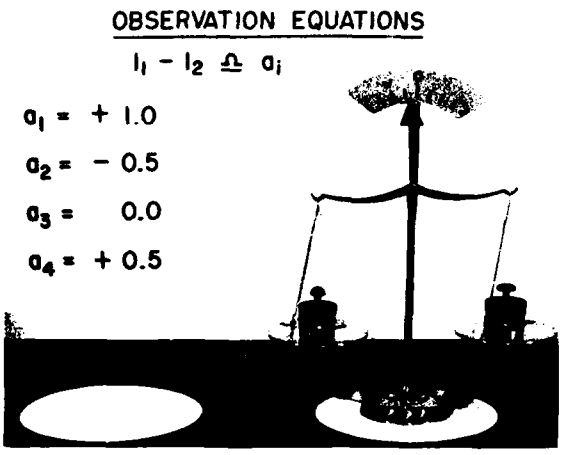


Figure 46

A characteristic of a measurement process is that it produces non-identical results. In our previous charts we had measurements of a 200 g weight, here are shown four measurements of the difference in mass. Through the redundancy—here 3 extra measurements—we get our grip on precision. In weight calibration we do not rely on repeated measurements of the same quantity but achieve the same result in another way.

★★★

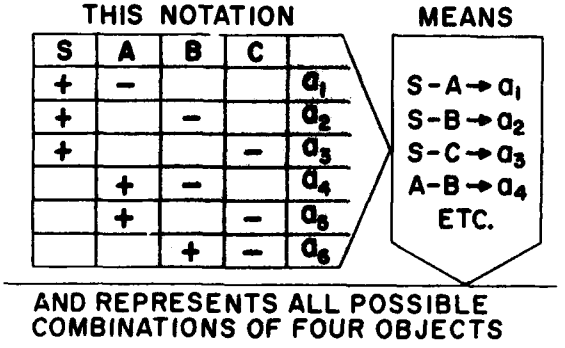


Figure 47

When we intercompare four objects, for example, four 1-kg standards, we could use six observations. Weight S is compared with A for a_1, S with B for a_2 and so on. If S were a standard and the rest unknowns, we again have 3 more measurements than we need and these serve to tell us of the precision of the process.

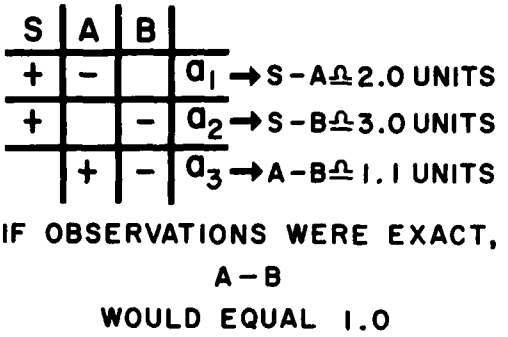


Figure 48

A simple example, using only three of the observations of the previous series, with S as the standard, A as the unknown, and B as the check standard, might give rise to the values shown. If everything were perfect, all equations representing the weighings would be satisfied exactly. Their lack of agreement would give a measure of the variability.

★★★

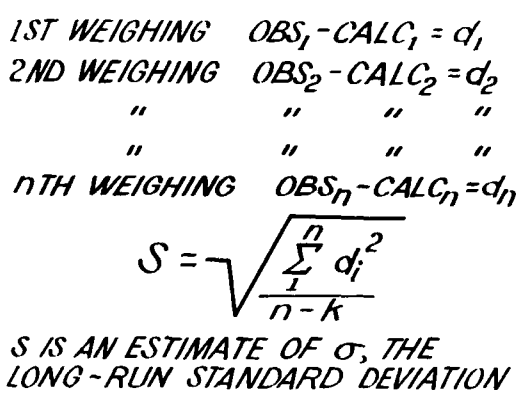


Figure 49

In general, for such weighing, there will be a discrepancy between the observed value and the best value calculated from the data, "best" meaning in most cases the value obtained in the method of least squares. If all is going well, none of these deviations will be too large, and also certain combinations of them, such as the sum of the squares, will also be well behaved. For statistical analysis the standard deviation, S, is used as the measure for describing variability. The quantity, S, is a function of the observational errors and will change with each set of data just as the values for the unknown weights do. (The quantity, k, is the number of unknowns in the system.)

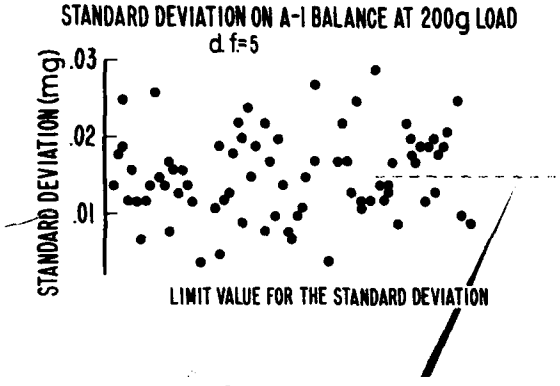


Figure 50

If the process is in a state of control these values of s will scatter about some value which is the true or long run standard deviation of the process.

★ ★ ★

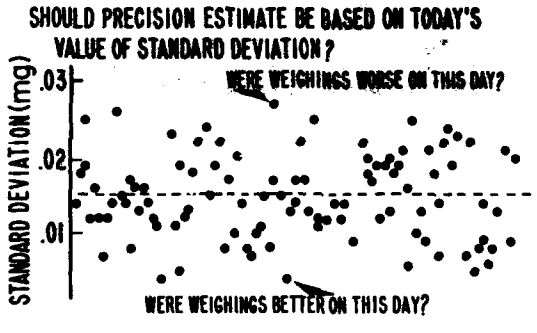


Figure 51

The argument that the uncertainty should be based on the internal agreement of today's values on the grounds that each day is unique or that weighing conditions are better on one day than on another may well be true. However, it will be expensive to make enough measurements on a given day to be sure that the variability has indeed changed from its long run average or to provide a reliable enough value to represent today's results. If the process did not change, using today's value would be analogous to keeping the last value of a sequence rather than using the mean represented by the dotted line. It is a sign that weighing conditions are not being reproduced, i.e., that the process is not in control, if the standard deviation does not stay within predicted limits. Let us now look again at the check standard.

Process Mean

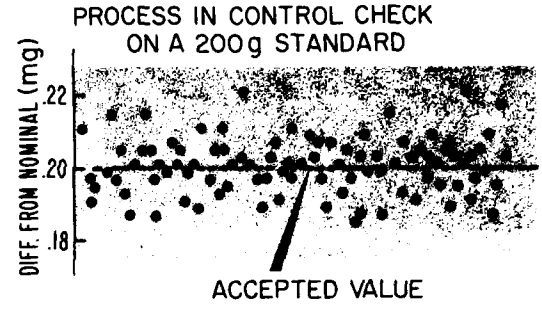


Figure 52

Each value obtained for the check standard serves not only a check on the process mean, but also can be used for evaluating the process variability. The same check standard, perhaps one of a group reserved for this purpose, is used consecutively in a given procedure until many independent values are obtained.

★ ★ ★

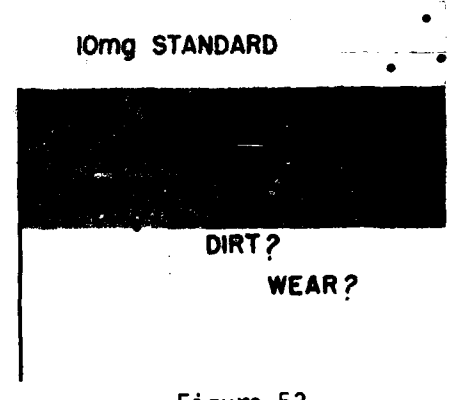


Figure 53

The importance of randomness cannot be overemphasized. As the collection of independent measurements on the check standard grows, it must be continually re-evaluated with reference to predicting the band within which the next point will lie. Slow drifts or sharp discontinuities are cause for concern until corrected, or satisfactorily explained.

Figure 54

If values return to normal after cleaning, one can rest easy, knowing the process is behaving properly. Indication of permanent changes are sometimes harder to explain, and even the most careful laboratories must occasionally repeat measurements because of troubles with foreign material adhering to or falling off the standard. If the new mean value persists over a sufficient number of measurements, it is proper to assume the standard has changed for some reason.

★★★

Process Control

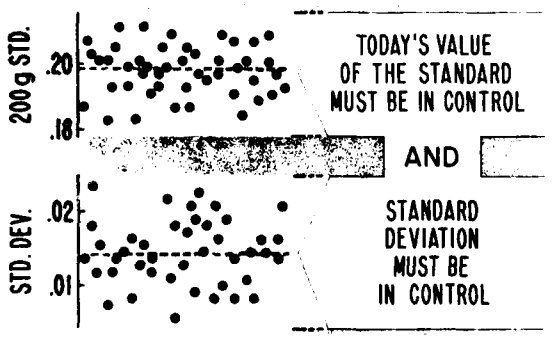


Figure 55

A check on *just* the value of the standard or *just* the precision is not enough. It turns out that the value for the precision and the value for the check standard are generally independent, that is, when s is small the deviation of the value determined for the check standard from the accepted value is equally often big and small. For control we need both conditions.

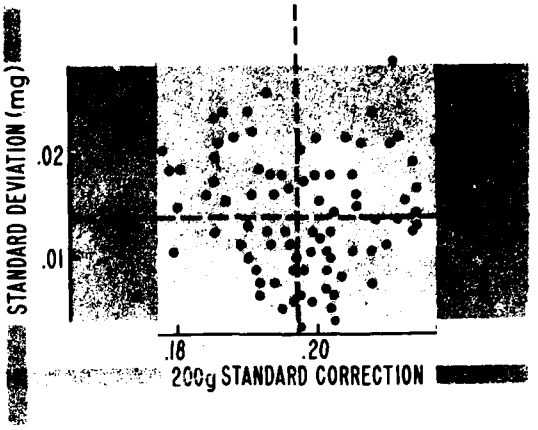


Figure 56

For a given set of observations the precision must be proper as shown on vertical scale and we must have a check on a known weight to establish that the limiting mean has not changed as shown on horizontal scale. Until these conditions are fulfilled, we cannot be sure exactly what it is that we are measuring. These are necessary conditions, and in perhaps most cases, also sufficient conditions to proclaim that the measurement process is in a state of control, as indicated by points within the central rectangle.

★★★

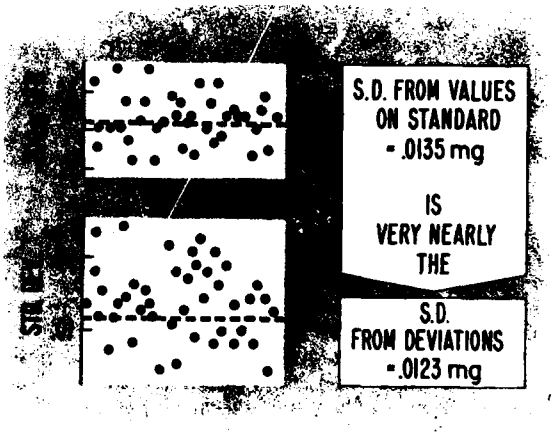


Figure 57

Because the check on the standard is spread over a considerable time interval, the variability will include the proper diversity of environmental and other factors and the sequence will, in the absence of seasonal or other systematic trouble, approximate

a sequence of independent values. If the weighing conditions are reproducible, then the daily standard deviation, s, and the variability as computed from the values of the check standard will be in agreement, i.e., the long run average of the variability as estimated from the control chart on the standard deviation should approach the corresponding value from the control chart based on the variability of the values of the check standard. Frequently, one is not in as good a shape as that indicated on the slide. When the measurements are spread out in time or space, an additional component of variation enters so that the lower chart gives an overly optimistic view of the process. A realistic estimate of process variability has to be based on that from the upper chart which reflects the total variation to which the measurements are subject. One would still use the within occasion variability for checking on control of the process, of course.

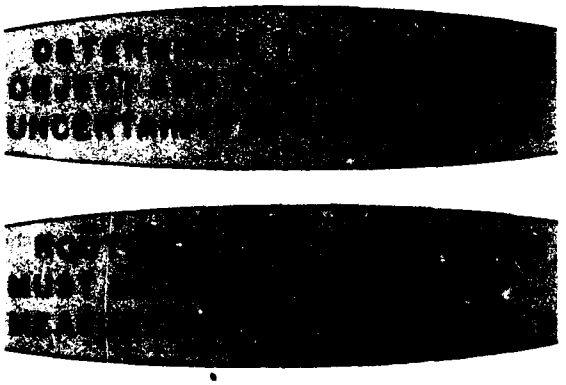


Figure 58

If in calibration we could measure the difference between the standard and the unknown again and again we could make an uncertainty statement similar to those just discussed for the case of measurements of a fixed difference, but in fact, we cannot routinely make enough measurements of this type to permit reliable estimates of the uncertainties.

★★★

Process Parameters and Uncertainty of Calibration

If we could be sure that our measurements of the difference between the unknown and the standard came from a process in a state of statistical control, that is to say a stable process with a known variability, then we could transfer the properties of the process to the individual measurement and be correct a stated percentage of the time.

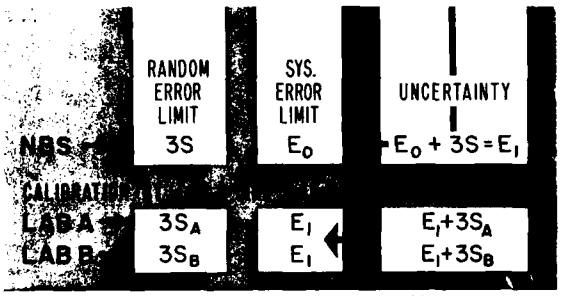
THE MEASUREMENT PROCESS REMAINS, AND IS, IN A SENSE, A CAPITAL INVESTMENT.

THE MEASUREMENTS, LIKE PRODUCTS, PASS ON TO OTHER DESTINATIONS.

Figure 59

All who weigh, or make other measurements, should concentrate on the properties of the measurement process—the degree to which the process re-creates the same value for its standards and exhibits the same level of variability. These are the properties that remain. The weights that are calibrated pass on to other destinations.

★★★



S, S_A AND S_B CAN BE NEARLY EQUAL. IF SO, THEN LAB A AND LAB B CAN CALIBRATE THIER OWN SET FROM SELECTED STANDARD WEIGHTS

Figure 60

At every stage in the extension of a measurement unit from an accepted standard to the ultimate user, there are three items of interest—a standard item, or items, with announced values and associated uncertainty, an assembly of equipment and procedures necessary for making the necessary comparisons, and the items which must be measured to accomplish some useful task. The uncertainty of the values established for the user are of paramount importance. This uncertainty has two components—one associated with the value of the starting standard and one reflecting the contribution of the local measurement process. The total uncertainty at any particular place becomes the systematic error for those who must use the service provided.

NATIONAL BUREAU OF STANDARDS
REPORT OF CALIBRATION TEST NO. 189160

INSTITUTE FOR BASIC STANDARDS
U.S. DEPARTMENT OF COMMERCE

FOR LOCKHEED-CALIFORNIA COMPANY
BURBANK, CALIFORNIA

ITEM SET OF MASS STANDARDS 10KG TO 1G DESIGNATED SET C
MFR. SEDERERER-KOHLBUSCH SERIAL NO. 781163
STATED DENSITY 10KG TO 1G 7.80 G PER CM3 AT 20C

THE ABOVE 1 ERS HAVE THE MASS VALUES SHOWN WITH REFERENCE TO THE NATIONAL STANDARD OF MASS. SEE ATTACHED SUPPLEMENT FOR LIMITATIONS IN USE OF APPARENT MASS VALUES AND UNCERTAINTY FROM [illegible] IN DENSITY.

ITEM	NOMINAL	APP MASS CORR (MG)	TRUE MASS CORR (MG)	UNCERTAINTY (MG)	VOL AT 20 (CM)3
1KG	1000.00	0.045700	9.126909	0.110227	26.74386883
500G	500.00	0.278974	4.819581	0.069271	63.37196684
300G	300.00	-0.265581	2.438779	0.063098	38.02312469
200G	200.00	-0.018969	1.797272	0.082346	25.34877014
100G	100.00	-0.093402	0.814718	0.064880	12.67437446
50G	50.00	-0.013024	0.441037	0.032909	6.32719146
30G	30.00	-0.000559	0.271878	0.020735	3.80231580
20G	2 .00	0.075933	0.257824	0.014434	2.53488687
10G	10.00	-0.119775	-0.027964	0.010710	1.26742237
5G	5.00	0.002880	0.048088	0.003708	0.63371865
3G	3.00	0.015487	0.042730	0.004076	0.38023355
2G	2.00	0.043916	0.062078	0.003082	0.25349329
1G	1.00	[illegible]	[illegible]	[illegible]	0.12673927

THE UNCERTAINTY FIGURE IS AN EXPRESSION OF THE OVERALL UNCERTAINTY USING THREE STANDARD DEVIATIONS AS A LIMIT TO THE EFFECT OF RANDOM ERRORS OF MEASUREMENT, THE MAGNITUDE OF SYSTEMATIC ERRORS FROM KNOWN SOURCES BEING NEGLIGIBLE.

FOR THE DIRECTOR,
PAUL E. PONTIUS, CHIEF

TEST COMPLETED AUGUST 17, 1966
WASHINGTON, D.C. 20234

MASS AND VOLUME SECTION

Figure 61

Any report of calibration or report of test must state a realistic uncertainty based on actual process performance. All of the pertinent data must be included so that the local processes can minimize the introduction of additional systematic errors. The random component of the uncertainty is a function of the measurement effort in the local process, reflecting the actual performance of that particular measurement process.

★★★

REPEATED CALIBRATION OF SAME OBJECT CLEARLY DEMONSTRATES THE PERFORMANCE OF A PARTICULAR MEASUREMENT PROCESS

Figure 62

There is no substitute for the evidence provided by the repeated calibration of the same object, over an extended time period, in demonstrating what the measurement process can do. These measurements should be independent repetitions, made under all the diversity of condition by which the method is affected so as to represent the set of conditions to which we wish our prediction to apply. The internally based precision estimate is applicable only to a narrower range of conditions, and it is only when the measurement conditions are highly reproducible that the two estimates of precision become equal.

★★★

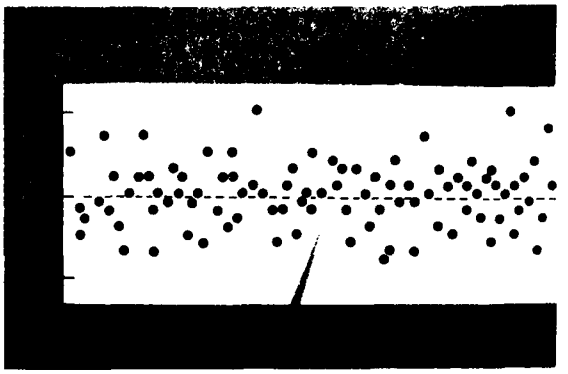


Figure 63

The routine calibration of one of the laboratory's weights, used as check standard, tells us what the process can do—it is not just a simulation of the calibration process—it is the real thing—without the need for any assumptions. It provides the basis for the precision statement or gives us a check on any internally based statement. We can say to our clients: "If we calibrate your weight a large number of times the results would look like those on the chart. We did it only once so that your value is like one of these points. Which one, we cannot say but we are fairly certain that it is within the indicated uncertainty."

★★★

JOURNAL OF RESEARCH of the National Bureau of Standards—C. Engineering and Instrumentation
Vol. 67C, No. 2, April-June 1963

Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems *

Churchill Eisenhart

(November 28, 1962)

Calibration of instruments and standards is a refined form of measurement. Measurement of some property of a thing is an operation that yields as an end result a number that indicates how much of the property the thing has. Measurement is ordinarily a repeatable operation, so that it is appropriate to regard measurement as a production process, the "product" being the numbers, i.e., the measurements, that it yields; and to apply to measurement processes in the laboratory the concepts and techniques of statistical process control that have proved so useful in the quality control of industrial production.

Viewed thus it becomes evident that a particular measurement operation cannot be regarded as constituting a measurement process unless statistical stability of the type known as a state of statistical control has been attained. In order to determine whether a particular measurement operation is, or is not, in a state of statistical control it is necessary to be definite on what variations of procedure, apparatus, environmental conditions, observers, operators, etc., are allowable in "repeated applications" of what will be considered to be the same measurement process applied to the measurement of the same quantity under the same conditions. To be realistic, the "allowable variations" must be of sufficient scope to bracket the circumstances likely to be met in practice. Furthermore, any experimental program that aims to determine the standard deviation of a measurement process as an indication of its precision, must be based on appropriate random sampling of this likely range of circumstances.

Ordinarily the accuracy of a measurement process may be characterized by giving (a) the standard deviation of the process and (b) credible bounds to its likely overall systematic error. Determination of credible bounds to the combined effect of recognized potential sources of systematic error always involves some arbitrariness, not only in the placing of reasonable bounds on the systematic error likely to be contributed by each particular assignable cause, but also in the manner in which these individual contributions are combined. Consequently, the "inaccuracy" of end results of measurement cannot be expressed by "confidence limits" corresponding to a definite numerical "confidence level," except in those rare instances in which the possible overall systematic error of a final result is negligible in comparison with its imprecision.

1. Introduction

Calibration of instruments and standards is basically a refined form of measurement. Measurement is the assignment of numbers to material things to represent the relations existing among them with respect to particular properties. One always measures properties of things, not the things themselves. In practice, measurement of some property of a thing ordinarily takes the form of a sequence of steps or operations that yields as an end result a number that indicates how much of this property the thing has, for someone to use for a specific purpose. The end result may be the outcome of a single reading of an instrument. More often it is some kind of average, e.g., the arithmetic mean of a number of independent determinations of the same magnitude, or the final result of a least squares "reduction" of measurements of a number of different quantities that bear known relations to each other in accordance with a definite experimental plan. In general, the purpose for which the answer is needed determines the accuracy required and ordinarily also the method of measurement employed.

Specification of the apparatus and auxiliary equipment to be used, the operations to be performed, the sequence in which they are to be executed, and the conditions under which they are respectively to be carried out—these instructions collectively serve to define a method of measurement. A measurement process is the realization of a method of measurement in terms of particular apparatus and equipment of the prescribed kinds, particular conditions that at best only approximate the conditions prescribed, and particular persons as operators and observers.

It has long been recognized that, in undertaking to apply a particular method of measurement, a degree of consistency among repeated measurements of a single quantity needs to be attained before the method of measurement concerned can be regarded as meaningfully realized, i.e., before a measurement process can be said to have been established that is

*Presented at the 1962 Standards Laboratory Conference, National Bureau of Standards, Boulder, Colo., August 8-10, 1962.

Reprinted with corrections, September 1968.

a realization of the method of measurement concerned. Indeed, consistency or statistical stability of a very special kind is required: to qualify as a measurement process a measurement operation must have attained what is known in industrial quality control language as a state of statistical control. Until a measurement operation has been "debugged" to the extent that it has attained a state of statistical control it cannot be regarded in any logical sense as measuring anything at all. And when it has attained a state of statistical control there may still remain the question of whether it is faithful to the method of measurement of which it is intended to be a realization.

The systematic error, or bias, of a measurement process refers to its tendency to measure something other than what was intended; and is determined by the magnitude of the difference $\mu-\tau$ between the process average or limiting mean μ associated with measurement of a particular quantity by the measurement process concerned and the true value τ of the magnitude of this quantity. On first thought, the "true value" of the magnitude of a particular quantity appears to be a simple straightforward concept. On careful analysis, however, it becomes evident that the "true value" of the magnitude of a quantity is intimately linked to the purposes for which knowledge of the magnitude of this quantity is needed, and cannot, in the final analysis, be meaningfully and usefully defined in isolation from these needs.

The precision of a measurement process refers to, and is determined by the degree of mutual agreement characteristic of independent measurements of a single quantity yielded by repeated applications of the process under specified conditions; and its accuracy refers to, and is determined by, the degree of agreement of such measurements with the true value of the magnitude of the quantity concerned. In brief "accuracy" has to do with closeness to the truth; "precision," only with closeness together.

Systematic error, precision, and accuracy are inherent characteristics of a measurement process and not of a particular measurement yielded by the process. We may also speak of the systematic error, precision, and accuracy of a particular method of measurement that has the capability of statistical control But these terms are not defined for a measurement operation that is not in a state of statistical control.

The precision, or more correctly, the *im*precision of a measurement process is ordinarily summarized by the standard deviation of the process, which expresses the characteristic disagreement of repeated measurements of a single quantity by the process concerned, and thus serves to indicate by how much a particular measurement is likely to differ from other values that the same measurement process might have provided in this instance, or might yield on remeasurement of the same quantity on another occasion. Unfortunately, there does not exist any single comprehensive measure of the accuracy (or *in*accuracy) of a measurement process analogous to the standard deviation as a measure of its imprecision.

To characterize the accuracy of a measurement process it is necessary, therefore, to indicate (a) its systematic error or bias, (b) its precision (or *im*precision)—and, strictly speaking, also, (c) the form of the distribution of the individual measurements about the process average. Such is the unavoidable situation if one is to concern one's self with individual measurements yielded by any particular measurement process. Fortunately, however, "final results" are ordinarily some kind of average or adjusted value derived from a set of independent measurements, and when four or more independent measurements are involved, such adjusted values tend to be normally distributed to a very good approximation, so that the accuracy of such final results can ordinarily be characterized satisfactorily by indicating (a) their imprecision as expressed by their standard error, and (b) the systematic error of the process by which they were obtained.

The error of any single measurement or adjusted value of a particular quantity is, by definition, the difference between the measurement or adjusted value concerned and the true value of the magnitude of this quantity. The error of any particular measurement or adjusted value is, therefore, a fixed number; and this number will ordinarily be unknown and unknowable, because the true value of the magnitude of the quantity concerned is ordinarily unknown and unknowable. Limits to the error of a single measurement or adjusted value may, however, be inferred from (a) the precision, and (b) bounds on the systematic error of the measurement process by which it was produced—but not without risk of being incorrect, because, quite apart from the inexactness with which bounds are commonly placed on a systematic error of a measurement process, such limits are applicable to the error of the single measurement or adjusted value, not as a unique individual outcome, but only as a typical case of the errors characteristic of such measurements of the same quantity that might have been, or might be, yielded by the same measurement process under the same conditions.

Since the precision of a measurement process is determined by the characteristic "closeness together" of successive independent measurements of a single magnitude generated by repeated application of the process under specified conditions, and its bias or systematic error is determined by the direction and amount by which such measurements tend to differ from the true value of the magnitude of the quantity concerned, it is necessary to be clear on what variations of procedure, apparatus, environmental conditions, observers, etc., are allowable in "repeated applications" or what will be considered to be the same measurement process applied to the measurement of the same quantity under the same conditions. If whatever measures of the precision and bias of a measurement process we may adopt are to provide a realistic indication of the accuracy of this process in practice, then the "allowable variations" must be of sufficient scope to bracket the range of circumstances commonly met in practice. Furthermore, any experimental program that aims to determine the pre-

cision, and thence the accuracy of a measurement process, must be based on an appropriate random sampling of this "range of circumstances," if the usual tools of statistical analysis are to be strictly applicable.

When adequate random sampling of the appropriate "range of circumstances" is not feasible, or even possible, then it is necessary (a) to compute, by extrapolation from available data, a more or less subjective estimate of the precision of the measurement process concerned, to serve as a substitute for a direct experimental measure of this characteristic, and (b) to assign more or less subjective bounds to the systematic error of the measurement process. To the extent that such at least partially subjective computations are involved, the resulting evaluation of the overall accuracy of a measurement process "is based on subject-matter knowledge and skill, general information, and intuition—but not on statistical methodology" [Cochran et al. 1953, p. 693]. Consequently, in such cases the statistically precise concept of a family of "confidence intervals" associated with a definite "confidence level" or "confidence coefficient" is not applicable.

The foregoing points and certain other related matters are discussed in greater detail in the succeeding sections, together with an indication of procedures for the realistic evaluation of precision and accuracy of established procedures for the calibration of instruments and standards that minimize as much as possible the subjective elements of such an evaluation. To the extent that complete elimination of the subjective element is not always possible, the responsibility for an important and sometimes the most difficult part of the evaluation is shifted from the shoulders of the statistician to the shoulders of the subject matter "expert."

2. Measurement

2.1. Nature and Object

Measurement is the assignment of numbers to material things to represent the relations existing among them with respect to particular properties. The number assigned to some particular property serves to represent the relative amount of this property associated with the object concerned.

Measurement always pertains to properties of things, not to the things themselves. Thus we cannot measure a meter bar, but can and usually do, measure its length; and we could also measure its mass, its density, and perhaps, also its hardness.

The object of measurement is twofold: first, symbolic representation of properties of things as a basis for conceptual analysis; and second, to effect the representation in a form amenable to the powerful tools of mathematical analysis. The decisive feature is symbolic representation of properties, for which end numerals are not the only usable symbols.

In practice the assignment of a numerical magnitude to a particular property of a thing is ordinarily accomplished by comparison with a set of standards, or by comparison either of the quantity itself, or of some transform of it, with a previously calibrated scale. Thus, length measurements are usually made by directly comparing the length concerned with a calibrated bar or tape; and mass measurements, by directly comparing the weight of a given mass with the weight of a set of standard masses, by means of a balance; but force measurements are usually carried out in terms of some transform, such as by reading on a calibrated scale the extension that the force produces in a spring, or the deflection that it produces in a proving ring; and temperature measurements are usually performed in terms of some transform, such as by reading on a calibrated scale the expansion of a column of mercury, or the electrical resistance of a platinum wire.

2.2. Qualitative and Quantitative Aspects

As Walter A. Shewhart, father of statistical control charts, has remarked:

> "It is important to realize . . . that there are two aspects of an operation of measurement; one is quantitative and the other qualitative. One consists of *numbers* or pointer readings such as the observed lengths in n measurements of the length of a line, and the other consists of the *physical manipulations* of physical things by *someone* in accord with instructions that we shall assume to be describable in words constituting a text." [Shewhart 1939, p. 130.]

More specifically, the qualitative factors involved in the measurement of a quantity are: the *apparatus* and *auxiliary equipment* (e.g., reagents, batteries or other source of electrical energy, etc.) employed; the *operators* and *observers*, if any, involved; the *operations* performed, together with the *sequence* in which, and the *conditions* under which, they are respectively carried out.

2.3. Correction and Adjustment of Observations

The numbers obtained as "readings" on a calibrated scale are ordinarily the end product of everyday measurement in the trades and in the home. In scientific work there are usually two important additional quantitative aspects of measurement: (1) *correction* of the readings, or their transforms, to compensate for known deviations from ideal execution of the prescribed operations, and for non-negligible effects of variations in uncontrolled variables; and (2) *adjustment* of "raw" or corrected measurements of particular quantities to obtain values of these quantities that conform to restrictions upon, or interrelations among, the magnitudes of these quantities imposed by the nature of the problem.

Thus, it may not be practicable or economically feasible to take readings at exactly the prescribed temperatures; but quite practicable and feasible to bring and hold the temperature within narrow neighborhoods of the prescribed values and to record the actual temperatures to which the respective readings correspond. In such cases, if the deviations from the prescribed temperatures are not negligible, "temperature *corrections*" based on appropriate theory are usually applied to the respective readings to bring

them to the values that presumably would have been observed if the temperature in each instance had been exactly as prescribed.

In practice, however, the objective just stated is rarely, if ever, actually achieved. Any "temperature corrections" applied could be expected to bring the respective readings "to the values that presumably would have been observed if the temperature in each instance had been exactly as prescribed" if and only if these "temperature corrections" made appropriate allowances for *all* of the effects of the deviations of the actual temperatures from those prescribed. "Temperature corrections" ordinarily correct *only* for particular effects of the deviations of the actual temperatures from their prescribed values; *not* for all of the effects on the readings traceable to deviations of the actual temperatures from those prescribed. Thus Michelson utilized *"temperature* corrections" in his 1879 investigation of the speed of light; but his results exhibit a dependence on temperature after "temperature correction." The "temperature corrections" applied corrected only for the effects of thermal expansion due to variations in temperature and not also for changes in the index of refraction of the air due to changes in the humidity of the air, which in June and July at Annapolis is highly correlated with temperature. *Corrections applied in practice are usually of more limited scope than the names that they are given appear to indicate.*

Adjustment of observations is fundamentally different from their "correction." When two or more related quantities are measured individually, the resulting measured values usually fail to satisfy the constraints on their magnitudes implied by the given interrelations among the quantities concerned. In such cases these "raw" measured values are mutually contradictory, and require *adjustment* in order to be usable for the purpose intended. Thus, measured values of the three cyclic differences $(A-B)$, $(B-C)$, and $(C-A)$ between the lengths of three nominally equivalent gage blocks are mutually contradictory, and strictly speaking are not usable as values of these differences, unless they sum to zero.

The primary goal of *adjustment* is to derive from such inconsistent measurements, if possible, *adjusted values* for the quantities concerned that do satisfy the constraints on their magnitudes imposed by the nature of the quantities themselves and by the existing interrelations among them. A second objective is to select from all possible sets of adjusted values the set that is the "best"—or, at least, a set that is "good enough" for the intended purpose—in some well-defined sense. Thus, in the above case of the measured differences between the lengths of three gage blocks, an adjustment could be effected by ignoring the measured value of one of the differences entirely, say, the difference $(C-A)$, and taking the negative of the sum of the other two as its adjusted value,

$$Adj(C-A)=-[(A-B)+(B-C)].$$

This will certainly assure that the sum of all three values, $(A-B)+(B-C)+Adj(C-A)$, is zero, as required, and is clearly equivalent to ascribing all of the excess or deficit to the replaced measurement, $(C-A)$. Alternatively, one might prefer to distribute the necessary total adjustment $-[(A-B)+(B-C)+(C-A)]$ equally over the individual measured differences, to obtain the following set of adjusted values:

$$Adj\,(A-B)=(A-B)-\frac{1}{3}[(A-B)+(B-C)+(C-A)]$$

$$=\frac{1}{3}[2(A-B)-(B-C)-(C-A)]$$

$$Adj\,(B-C)=\frac{1}{3}[2(B-C)-(A-B)-(C-A)]$$

$$Adj\,(C-A)=\frac{1}{3}[2(C-A)-(A-B)-(B-C)]$$

Clearly, the sum of these three adjusted values must always be zero, as required, regardless of the values of the original individual measured differences. Furthermore, most persons, I believe, would consider this latter adjustment the better; and under certain conditions with respect to the "law of error" governing the original measured differences, it is indeed the "best."

Note that no adjustment problem existed at the stage when only two of these differences had been measured whichever they were, for then the third could be obtained by subtraction. As a general principle, when no more observations are taken than are sufficient to provide one value of each of the unknown quantities involved, then the results so obtained are usable at least—they may not be "best." On the other hand, when additional observations are taken, leading to "over determination" and consequent contradiction of the fundamental properties of, or the basic relationships among the quantities concerned, then the respective observations must be regarded as contradicting one another. When this happens the observations themselves, or values derived from them, must be replaced by adjusted values such that all contradiction is removed. "This is a logical necessity, since we cannot accept for truth that which is contradictory or leads to contradictory results." [Chauvenet 1868, p. 472.]

2.4. Scheduling the Taking of Measurements

Having done what one can to remove extraneous sources of error, and to make the basic measurements as precise and as free from systematic error as possible, it is frequently possible not only to increase the precision of the end results of major interest but also to simultaneously decrease their sensitivity to sources of possible systematic error, by careful scheduling of the measurements required. An instance is provided by the traditional procedure for calibrating liquid-in-glass thermometers [Waidner and Dickinson 1907, p. 702; NPL 1957, pp. 29–30; Swindells 1959, pp. 11–12]: Instead of attempting to hold the temperature of the comparison bath constant, a very difficult objective to achieve, the heat

input to the bath is so adjusted that its temperature is slowly increasing at a steady rate, and then readings of, say, four test thermometers and two standards are taken in accordance with the schedule

$$S_1T_1T_2T_3T_4S_2S_2T_4T_3T_2T_1S_1$$

the readings being spaced uniformly in time so that the arithmetic mean of the two readings of any one thermometer will correspond to the temperature of the comparison bath at the midpoint of the period. Such scheduling of measurement taking operations so that the effects of the specific types of departures from perfect control of conditions and procedure will have an opportunity to balance out is one of the principal aims of the art and science of *statistical design of experiments*. For additional physical science examples, see, for instance, Youden [1951a; and 1954–1959].

2.5. Measurement as a Production Process

We may summarize our discussion of measurement up to this point, as follows: Measurement of some property of a thing in practice always takes the form of a sequence of steps or operations that yield as an end result a number that serves to represent the amount or quantity of some particular property of a thing—a number that indicates how much of this property the thing has, for someone to use for a specific purpose. The end result may be the outcome of a single reading of an instrument, with or without corrections for departures from prescribed conditions. More often it is some kind of average or adjusted value, e.g., the arithmetic mean of a number of independent determinations of the same magnitude, or the final result of, say, a least squares "reduction" of measurements of a number of different quantities that have known relations to the quantity of interest.

Measurement of some property of a thing is ordinarily a repeatable operation. This is certainly the case for the types of measurement ordinarily met in the calibration of standards and instruments. It is instructive, therefore, to regard measurement as a *production process*, the "product" being the numbers, that is, the measurements that it yields; and to compare and contrast measurement processes in the laboratory with mass production processes in industry. For the moment it will suffice to note (a) that when successive amounts of units of "raw material" are processed by a particular mass production process, the output is a series of nominally identical items of product—of the particular type produced by the mass production operation, i.e., by the *method of production* concerned; and (b) that when successive objects are measured by a particular measurement process, the individual items of "product" produced consist of the numbers assigned to the respective objects to represent the relative amounts that they possess of the property determined by the *method of measurement* involved

2.6. Methods of Measurement and Measurement Processes

Specification of the apparatus and auxiliary equipment to be used, the operations to be performed, the sequence in which they are to be carried out, and the conditions under which they are respectively to be carried out—these *instructions* collectively serve to define a *method of measurement*. To the extent that corrections may be required they are an integral part of measurement. The types of corrections that will ordinarily need to be made, and specific procedures for making them, should be included among "the operations to be performed." Likewise, the essential adjustments required should be noted, and specific procedures for making them incorporated in the specification of a method of measurement.

A *measurement process* is the realization of a method of measurement in terms of particular apparatus and equipment of the prescribed kinds, particular conditions that at best only approximate the conditions prescribed, and particular persons as operators and observers [ASTM 1961, p. 1758; Murphy 1961, p. 264]. Of course, there will often be a question whether a particular measurement process is loyal to the method of measurement of which it is intended to be a realization; or whether two different measurement processes can be considered to be realizations of the same method of measurement.

To begin with, written specifications of methods of measurement often contain absolutely precise instructions which, however, cannot be carried out (repeatedly) with complete exactitude in practice; for example, "move the two parallel cross hairs of the micrometer of the microscope until the graduation line of the standard is centered between them." The accuracy with which such instructions can be carried out in practice will always depend upon "the circumstances"; in the case cited, on the skill of the operator, the quality of the graduation line of the standard, the quality of the screw of the micrometer the parallelism of the cross hairs, etc. To the extent that the written specification of a method of measurement involves absolutely precise instructions that cannot be carried out with complete exactitude in practice there are certain to be discrepancies between a method of measurement and its realization by a particular measurement process.

In addition, the specification of a method of measurement often includes a number of imprecise instructions, such as "raise the temperature slowly," "stir well before taking a reading," "make sure that the tubing is clean," etc. Not only are such instructions inherently vague, but also in any given instance they must be understood in terms of the general level of refinement characteristic of the context in which they occur. Thus, "make sure that the tubing is clean" is not an absolutely definite instruction; to some people this would mean simply that the tubing should be clean enough to drink liquids through; in some laboratory work it might be interpreted to mean mechanically washed and scoured so as to be free from dirt and other ordinary

solid matter (but not cleansed also with chemical solvents to remove more stubborn contaminants); to an advanced experimental physicist it may mean not merely mechanically washed and chemically cleansed, but also "out gassed" by being heated to and held at a high temperature, near the softening point, for an hour or so. All will agree, I believe, that it would be exceedingly difficult to make such instructions absolutely definite with a convenient number of words. To the extent that the specification of a method of measurement includes instructions that are not absolutely definite, there will be room for differences between measurement processes that are intended to be realization of the very same method of measurement.

Recognition of the difficulty of achieving absolute definiteness in the specification of a method of measurement does not imply that "any old set" of instructions will serve to define a method of measurement. Quite the contrary. To qualify as a specification of a method of measurement, a set of instructions must be sufficiently definite to insure statistical stability of repeated measurements of a single quantity, that is, derived measurement processes must be capable of meeting the criteria of *statistical control* [Shewhart 1939, p. 131; Murphy 1961, p. 265; ASTM 1961, p. 1758]. To elucidation of the meaning of, and need for this requirement we now turn.

3. Properties of Measurement Processes

3.1. Requirement of Statistical Control

The need for attaining a degree of consistency among repeated measurements of a single quantity before the method of measurement concerned can be regarded as meaningful has certainly been recognized for a long, long time. Thus Galileo, describing his famous experiment on the acceleration of gravity in which he allowed a ball to roll different distances down an inclined plane wrote:

". . . si lasciava (come dico) scendere per il detto canale la palla, notando, nel modo che appresso dirò, il tempo che consumava nello scorrerlo tutto, replicando il medesimo atto molte volte per assicurarsi bene della quantità del tempo, nel quale non si trovava mai differenza nè anco della decima parte d'una battuta di polso. Fatta e stabilita precisamente tale operazione, facemmo scender la medisima palla solamente per la quarta parte della lunghezza di esso canale . . ."[1] [Galileo 1638, Third Day; Nat'l. ed., p. 213.]

Something more than mere "consistency" is required, however, as Shewhart points out eloquently in his very important chapter on "The Specification of Accuracy and Precision" [Shewhart 1939, ch. IV]. He begins by noting that the description given by R. A. Millikan [1903, pp. 195–196] of a method for determining the surface tension T of a liquid from measurements of the force of tension F of a film of the liquid contains the following instruction with regard to the basic readings from which measurements of F are derived: "Continue this operation until a number of consistent readings can be obtained." Shewhart then comments on this as follows:

". . . the text describing the operation does not say to carry out such and such physical operations and call the result a measurement of T. Instead, it says in effect not to call the result a measurement of T until one has attained a certain degree of *consistency* among the observed values of F and hence among those of T. Although this requirement is not always explicitly stated in specifications of the operation of measurements as it was here, I think it is always implied. Likewise, I think it is always assumed that there can be too much consistency or uniformity among the observed values as, for example, if a large number of measurements of the surface tension of a liquid were found to be identical. What is wanted but not explicitly described is a specific kind and degree of consistency.

". . . it should be noted that the advice to repeat the operation of measuring surface tension until a number of consistent readings have been obtained is indefinite in that it does not indicate how many readings shall be taken before applying a test for consistency, nor what kind of test of consistency is to be applied to the numbers or pointer readings . . . . One of the objects of this chapter is to see how far one can go toward improving this situation by providing an operationally definite criterion that preliminary observations must meet before they are to be considered consistent in the sense implied in the instruction cited above.

"Before doing this, however, we must give attention not so much to the consistency of the n observed values already obtained by n repetitions of the operation of measurement as we do to the *reproducibility of the operation* as determined by the numbers in the potentially infinite sequence corresponding to an infinite number of repetitions of this operation. No one would care very much how consistent the first n preliminary observations were if nothing could be validly inferred from this as to what future observations would show. Hence, it seems to me that the characteristics of the numerical aspects of an operation that is of greatest practical interest is its *reproducibility within tolerance limits throughout the infinite sequence*. The limit to which we may go in this direction is to attain a state of statistical control. The attempt to attain a certain kind of consistency within the first n observed values is merely a means of attaining reproducibility within limits throughout the whole of the sequence." [Shewhart 1939, pp. 131–132.]

The point that Shewhart makes forcefully, and stresses repeatedly later in the same chapter, is that the first n measurements of a given quantity generated by a particular measurement process provide a logical basis for predicting the behavior of further measurements of the same quantity by the same measurement process if and only if these n measurements may be regarded as a *random sample* from a "population" or "universe" of all conceivable measurements of the given quantity by the measurement process concerned; that is, in the language of mathematical statistics, if an only if the n measurements in hand may be regarded as "observed values" of a sequence of random variables characterized by a probability distribution identified with the measurement process concerned, and related through the values of one or more of its parameters to the magnitude of the quantity measured.

It should be noted especially that nothing is said about the mathematical form of the *probability distribution* of these random variables. The important thing is that there be one. W. Edwards

[1] I am grateful to my colleague Ugo Fano for the following literal translation: ". . . we let, as I was saying, the ball descend through said channel, recording, in a manner presently to be described, the time it took in traversing it all, repeating the same action many times to make really sure of the magnitude of time, in which one never found a difference of even a tenth of a pulsebeat. Having done and established precisely such operation, we let the same ball descend only for the fourth part of the length of the same channel; . . ."

Deming has put this clearly and forcefully in these words:

"In applying statistical theory, the main consideration is not what the shape of the universe is, but whether there is any universe at all. No universe can be assumed, nor . . . statistical theory . . . applied unless the observations show statistical control. In this state the samples when cumulated over a suitable interval of time give a distribution of a particular shape, and this shape is reproduced hour after hour, day after day, so long as the process remains in statistical control i.e., exhibits the properties of randomness. In a state of control, n observations may be regarded as a sample from the universe of whatever shape it is. A big enough sample, or enough small samples, enables the statistician to make meaningful and useful predictions about future samples. This is as much as statistical theory can do.

". . . Very often the experimenter, instead of rushing in to apply [statistical methods] should be more concerned about attaining statistical control and asking himself whether any predictions at all (the only purpose of his experiment), by statistical theory or otherwise, can be made." [Deming 1950, pp. 502–503.]

Shewhart was well aware of the fact that from a set of n measurements in hand it is not possible to decide with absolute certainty whether they do or do not constitute a *random sample* from some definite statistical "population" characterized by a probability distribution. He, therefore, proposed [Shewhart 1939, pp. 146–147] that in any particular instance one should "decide to act for the present as if"[2] the measurements in hand (and their immediate successors) were a simple random sample from a definite statistical population—i.e., in the language of mathematical statistics, were "observed values" of *independent identically distributed random variables* only if the measurements in hand met the requirements of the small-samples version of Criterion I of his previous book [Shewhart 1931, pp. 309–318] and of certain additional tests of randomness that he described explicitly for the first time in his contribution to the University of Pennsylvania Bicentennial Conference in September 1940 [Shewhart, 1941]. In other words, Shewhart proposed that one should consider a measurement process to be—i.e., should "decide to act for the present as if" the process were—in a *state of (simple) statistical control*, only if the measurements in hand show no evidence of lack of statistical control when analyzed for randomness *in the order in which they were taken* by the control chart techniques for averages and standard deviations that he had found so valuable in industrial process control and by certain additional tests for randomness based on "runs above and below average" and "runs up and down."[3]

Simpson[3] did not prove that taking of the Arithmetic Mean was the best thing to do but merely that it is good. However, in accomplishing this goal he did something much more important: he took the bold step of regarding errors of measurement, not as unique unrelated magnitudes unamenable to mathematical analysis, but as distributed in accordance with a probability distribution that was an intrinsic property of the measurement process itself. He thus opened the way to a mathematical theory of measurement based on the mathematical theory of probability; and, in particular, to the formulation and development of the Method of Least Squares in essentially its present day form by Gauss (1809, 1821) and Laplace (1812).

"Student" (William Sealy Gosset, 1876–1937), pioneer statistical consultant and "father" of the "theory of small samples," was certainly among the first to stress the importance of randomness in measurement and experimentation. Thus, he began his revolutionary 1908 paper on "The probable error of a mean" with these remarks:

"Any experiment may be regarded as forming an individual of a 'population' of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

"Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong." [Student 1908, p. 1.]

None of these writers, nor any of their contemporaries, however, provided "an operationally definite criterion that preliminary observations must meet" before we take it upon ourselves "to act for the present as if" they and their immediate successors were random samples from a "population" or "universe" of all conceivable measurements of the given quantity by the measurement process concerned. Provision of such a criterion is Shewhart's major contribution.

Experience shows that in the case of measurement processes the ideal of strict statistical control that Shewhart prescribes is usually very difficult to attain, just as in the case of industrial production processes. Indeed, many measurement processes simply do not and, it would seem, cannot be made to conform to this ideal of producing successive measurements of a single quantity that can be considered to be "observed values" of independent identically distributed random variables.[4] The nature of the "trouble" was stated succinctly by Student in 1917 when, speaking of physical and chemical determinations, he wrote:

"After considerable experience I have not encountered any determination which is not influenced by the date on which it is made; from this it follows that a number of determinations of the same thing made on the same day are likely

---

[2] This very explicit phraseology is due to John W. Tukey [1960, p. 424].

[3] Thomas Simpson, in his now famous letter [Simpson 1755] to the President of the Royal Society of London "on the Advantage of taking the Mean of a Number of Observations, in practical Astronomy," was the first to consider repeated measurements of a single quantity by a given measurement process as observed values of independent random variables having the same probability distribution. His conclusion is of interest in itself:

"Upon the whole of which it appears, that the taking of the Mean of a number of observations, greatly diminishes the chances for all the smaller errors, and cuts off almost all possibility of any great ones: which last consideration, alone, seem sufficient to recommend the use of the method, not only to astronomers, but to all others concerned in making of experiments of any kind (to which the above reasoning is equally applicable). And the more observations or experiments there are made, the less will the conclusion be liable to err, provided they admit of being repeated under the same circumstances."

[4] Looking at the matter from a fundamental viewpoint, perhaps we should say, not that Shewhart's ideal of strict statistical control is unattainable in the case of such measurement processes, but rather that the degree of approximation to this ideal can be made as close as one chooses, if one is willing to pay the price. In other words, how close one chooses to bring a measurement process to the ideal of strict statistical control is, in any given instance, basically an economic matter, taking into account, of course, not only the immediate purpose(s) for which the measurements are intended but also the other uses to which they may be put. (Compare Simon [1946, p. 566] and Eisenhart [1952, p. 554].)

to lie more closely together than if the repetitions had been made on different days." [Student 1917, p. 415.]

In other words, production of measurements seems to be like the production of paint; and just as in the case of paint, if one must cover a large surface all of which is visible simultaneously, one will do well to use paint all from the same batch, so in the case of measurements, if a scientist or metrologist "wishes to impress his clients" he will "arrange to do repetition analyses as nearly as possible at the same time." [Student 1927, p. 155.]

Fortunately, just as one may blend paint from several batches to obtain a more uniform color, and one which is, presumably, closer to the "process average," so also may a scientist or metrologist "if he wishes to diminish his real error, . . . separate [his measurements] by as wide an interval of time as possible" [Student, loc. cit.] and then take an appropriate average of them as his determination. Consequently, if we are to permit such averaging as an allowable step in a fully specified measurement process (see sec. 2.6 above), then we are obliged to recognize both within-day and between-day components of variation, and accept such a complex measurement process as being in a state of statistical control overall, or as we shall say, in a *state of COMPLEX statistical control*, when the components of within-day and between-day variation are both in a state of statistical control in Shewhart's strict sense, which we shall term *SIMPLE statistical control*. In more complex situations, one may be obliged to recognize more than two "layers" of variation, and, sometimes, more than a single component of variation within a given "layer."

Adopting this more general concept of statistical control, R. B. Murphy of the Bell Telephone Laboratories in his essay "On the Meaning of Precision and Accuracy" [Murphy 1961], published in advance of the issuance by the American Society for Testing and Materials of its Tentative Recommended Practice with respect to the "Use of the Terms Precision and Accuracy as Applied to Measurement of a Property of a Material" [ASTM 1961], remarks:

"Following through with this line of thought borrowed from quality control, we shall add a requirement that an effort to follow a test method ought not to be known as a measurement process unless it is capable of statistical control. Capability of control means that either the measurements are the product of an identifiable statistical universe or an orderly array of such universes or, if not, the physical causes preventing such identification may themselves be identified and, if desired, isolated and suppressed. Incapability of control implies that the results of measurement are not to be trusted as indications of the physical property at hand—in short, we are not in any verifiable sense measuring anything . . . . Without this limitation on the notion of measurement process, one is unable to go on to give meaning to those statistical measures which are basic to any discussion of precision and accuracy." [Murphy 1961, pp. 264–265.]

3.2. Postulate of Measurement and the Concept of a Limiting Mean

A conspicuous characteristic of measurement is disagreement of repeated measurements of the same quantity. Experience shows that, when high accuracy is sought, repeated measurements of the same quantity by a particular measurement process does not yield uniformly the same number.[5] We explain these discordances by saying that the individual measurements are affected by *errors*, which we interpret to be the manifestations of variations in the execution of the process of measurement resulting from "the imperfections of instruments, and of organs of sense," and from the difficulty of achieving (or even specifying with a convenient number of words) the ideal of perfect control of conditions and procedure.

This "cussedness of measurements" brings us face to face with a fundamental question: In what sense can we say that the measurements yielded by a particular measurement process serve to determine a unique magnitude, when experience shows that repeated measurement of a single quantity by this process yields a sequence of nonidentical numbers. What is the value thus determined?

The answer takes the form of a *postulate* about measurement processes that has been expressed by N. Ernest Dorsey, as follows:

"The mean of a family of measurements—of a number of measurements for a given quantity carried out by the same apparatus, procedure and observer—approaches a definite value as the number of measurements is indefinitely increased. Otherwise, they could not properly be called measurements of a given quantity. In the theory of errors, this limiting mean is frequently called the 'true' value, although it bears no necessary relation to the true quaesitum, to the actual value of the quantity that the observer desires to measure. This has often confused the unwary. Let us call it the limiting mean." [Dorsey 1944, p. 4, Dorsey and Eisenhart 1953, p. 103.]

In my lectures at the National Bureau of Standards, and elsewhere, I have termed this—or rather a slightly rephrased version of it—the *Postulate of Measurement*. A mathematical basis for it is provided by the Strong Law of Large Numbers, a theorem in the mathematical theory of probability discovered during the present century. See, for example, Feller [1957, pp. 243–245, 374], Gnedenko [1962, pp. 241–249], or Parzen [1960, p. 420].

Needless to say, by a "family of measurements" Dorsey means, not a succession of "raw" readings, but rather a succession of adjusted or corrected values which, by virtue of adjustment or correction, can rightfully be considered to be determinations of a single magnitude.

a. Mathematical Formulation

The foregoing can be expressed mathematically as follows: on some particular occasion, say the ith, we may take a number of successive measurements of a single quantity by a given measurement process under certain specified circumstances. Let

$$x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots \qquad (1)$$

[5] The qualification "when high accuracy is sought" is essential; for if using an ordinary two-pan chemical balance we measure and record the mass of a small metallic object only to the nearest gram, then we would expect all of our measurements to be the same—except in the equivocal case of a mass equal, or very nearly equal, to an odd multiple of $\frac{1}{2}$ g, and such equivocal cases can be resolved easily by adding a $\frac{1}{2}$ g mass to one pan. Full accordance of measurements clearly cannot be taken as incontestable evidence of high accuracy; but rather should be regarded as evidence of limited accuracy.

denote the sequence of measurements so generated. Conceptually at least, this sequence could be continued indefinitely. Likewise, on different occasions we might start a new sequence, using the same measurement procedure and applying it to measurement of the same quantity under the same fixed set of circumstances. Each such fresh "start" would correspond to a different value of i. If, for example, the measurement process concerned is statistically stable in the sense of being in a *state of statistical control* as defined by Shewhart [1939], then the Strong Law of Large Numbers will be applicable and we may expect the sequence of cumulative arithmetic means on the ith occasion, namely,

$$\bar{x}_{in} \equiv (x_{i1} + x_{i2} + \ldots + x_{in})/n, \quad (n = 1, 2, \ldots), \quad (2)$$

to converge to μ, a number that constitutes the limiting mean associated with the quantity measured by this measurement process under the circumstances concerned, but independent of the "occasion," that is, independent of the value of "i." The Strong Law of Large Numbers does not guarantee that the sequence (2) for a particular value of "i" will converge to μ as the number of observations n on this occasion tends to infinity, but simply states that among the family of such sequences corresponding to a large number of different starts, ($i = 1, 2, \ldots$), the *instances of nonconvergence to* μ will be *rare exceptions*. In other words, if the measurement process with which one is concerned satisfies the conditions for validity of the Strong Law of Large Numbers, then in practice one is almost certain to be working with a "good" sequence—one for which (2) would converge to μ if the number of observations were continued indefinitely—but "bad" occasions *can* occur, though rarely. Thus, the Postulate of Measurement expresses something better than an "on-the-average" property—it expresses an "in-almost-all-cases" property. Furthermore, this limiting mean μ, the value of which each individual measurement x is trying to express, can be regarded not only as the *mean* or "*center of gravity*" of the infinite conceptual population of all measurements x that might conceivably be generated by the measurement process concerned under the specified circumstances, but also as *the* value of the quantity concerned as determined by *this* measurement process.

b. Aim of the Postulate

The sole aim of the Postulate of Measurement is axiomatic acceptance of the existence of a limit approached by the arithmetic mean of a finite number n of measurements generated by any measurement process as $n \to \infty$. It says nothing about how the "best" estimate of this limiting mean is to be obtained from a finite number of such observations. The Postulate is an answer to the need of the practical man for a justification of his desire to consider the sequence of nonidentical numbers that he obtains when he attempts to measure a quantity "by the *same method* under like circumstances" as pertaining to a single magnitude, in spite of the evident discordance of its elements. The Postulate aims to satisfy this need by telling him that if he were to continue taking more and still more measurements on this quantity "by the same method under like circumstances" ad infinitum, and were to calculate their cumulative arithmetic means at successive stages of this undertaking, then he would find that the successive terms of this sequence of cumulative arithmetic means would settle down to a narrower and ever narrower neighborhood of some definite number which he could then accept as *the* value of the magnitude that his first few measurements were striving to express.

c. Importance of Limiting Mean

The concept of a *limiting mean* associated with the measurement of a given quantity by a particular measurement process that is in a *state of statistical control* is important because by means of statistical methods based on the mathematical theory of probability we can make quantitative inferential statements, with known chances of error, about the magnitude of this limiting mean from a set of measurements of the given quantity by the measurement process concerned. The magnitude of the limiting mean associated with the measurement of a given quantity by a particular measurement process must be carefully distinguished from the *true magnitude* of the quantity measured, about which we may be tempted to make similar inferential statements. Insofar as we make statistical inferences from a set of measurements, we make them with respect to a property of the measurement process involved under the circumstances concerned. The step from quantitative inferential statements about the limiting mean associated with the measurement of a given quantity by a particular measurement process, to quantitative statements about the true magnitude of the quantity concerned, may be based on subject matter knowledge and skill, general information and intuition—but not on statistical methodology. (Compare Cochran, Mosteller, and Tukey [1953, pp. 692–693].)

3.3. Definition of the Error of a Measurement, and of the Systematic Error, Precision, and Accuracy of a Measurement Process

a. Error of a Single Measurement or Adjusted Value

The *error* of any measurement of a particular quantity is, by definition, the difference between the measurement concerned and the *true value* of the magnitude of this quantity, taken positive or negative accordingly as the measurement is greater or less than the true value. In other words, if x denotes a single measurement of a quantity, or an adjusted value derived from a specific set of individual measurements, and τ is the *true value* of the magnitude of the quantity concerned, then, by definition,

$$\text{the } \textit{error} \text{ of } x \text{ as a measurement of } \tau \equiv x - \tau.$$

The error of any particular measurement or adjusted value, x, is, therefore, a fixed number. The

numerical magnitude and sign of this number will ordinarily be unknown and unknowable, because the true value of the magnitude of the quantity concerned is ordinarily unknown and unknowable. Limits to the error of a single measurement or adjusted value may, however, be inferred from (a) the *precision*, and (b) bounds on the *systematic error*, of the measurement process by which it was produced - but not without risk of being incorrect, because, quite apart from the inexactness with which bounds are commonly placed on the systematic error of a measurement process, such limits are applicable to the error of a single measurement or adjusted value, not as a unique individual outcome, but only as a typical case of the errors characteristic of measurements of the same quantity that might have been, or might be, yielded by the same measurement process under the same conditions.

b. Systematic Error of a Measurement Process

When the limiting mean μ associated with measurement of the magnitude of a quantity by a particular measurement process does not agree with the *true value* τ of the magnitude concerned, the measurement process is said to have a *systematic error*, or *bias*, of magnitude $\mu-\tau$.

The systematic error of a measurement process will ordinarily have both constant and variable components. Consider, for example, measurement of the distance between two points by means of a graduated metal tape [Holman 1892, p. 9]. Possible causes of systematic error that immediately come to mind are:

(1) Mistakes in numbering the scale divisions of the tape;
(2) irregular spacing of the divisions of the tape;
(3) sag of tape;
(4) stretch of tape;
(5) temperature not that for which the tape was calibrated.

For *any single distance*, the effects of (1) and (2) will be constant; and the effects of (3) and (4) will undoubtedly each contain a constant component characteristic of the distance concerned. Some of these effects will be of one sign, some of the other, and their algebraic sum will determine the *constant error* of this measurement process with respect to the particular distance concerned. Furthermore, the "constant error" of this measurement process will be different (at least, conceptually) for different distances measured.

In the case of repeated measurement of a single distance, the effect of (5), and at least portions of the effects of (3) and (4), may be expected to vary from one "occasion" to the next (e.g., from day to day), thus contributing *variable components* to the *systematic error* of the process.

A large fraction of the variable contributions of (3) and (4) could, and in practice no doubt would, be removed by stretching the tape by a spring balance or other means so that it is always under the same tension. The stretch corresponding to a particular distance would then be nearly the same at all times, and a fixed correction could be made for most of the sag corresponding to this distance. Furthermore, the effect of (5) could, and in practice probably would, be reduced by determining the temperature of the tape at various points along its length and applying a temperature correction. By comparison of the tape with a standard, the error arising from (1) could be eliminated entirely, and corrections determined as a basis for eliminating, or at least, reducing the effect of (2).

As in the foregoing example there are usually certain obvious sources of systematic error. Unfortunately, there are generally additional sources of systematic error, the detection, diagnosis, and eradication of which call for much patience and acumen on the part of the observer. The work involved in their detection, diagnosis, and eradication often far exceeds that of taking the final measurements, and is sometimes discouraging to the experienced observer as well as to the beginner. Fortunately, there are various statistical tools that are helpful in this connection, and Olmstead [1952] has found that of these the two most effective and universally useful are the average ($\bar{x}$) and range (R) charts of industrial quality control. (For details on the construction and use of $\bar{x}$- and R-charts, see, for example, the ASTM Manual on Quality Control of Materials [ASTM 1951, pp. 61-63 and p. 83]; or American Standards Z1.2-1958 and Z1.3-1958 [ASA 1958b, ASA 1958c].)

c. Concept of True Value

In the foregoing we have defined the *error* of a measurement x to be the difference $x-\tau$ between the measurement and the *true value* τ of the magnitude of the quantity concerned; and the *systematic error*, or *bias*, of a measurement process as the difference $\mu-\tau$ between the limiting mean μ associated with the measurement of a particular quantity by the measurement process concerned, and the *true value* τ of the magnitude of this quantity. This immediately raises the question: Just how is the "true value" of the magnitude of a particular property of some thing defined? In the final analysis, the "true value" of the magnitude of a quantity is defined by agreement among experts on an *exemplar method* for the measurement of its magnitude - it is the limiting mean of a conceptual *exemplar process* that is an ideal realization of the agreed-upon exemplar method. And the refinement to which one should go in specifying the exemplar process will depend on the purposes for which a determination of the magnitude of the quantity concerned is needed—not just the immediate purpose for which measurements are to be taken but also the other uses to which these measurements, or a final adjusted value derived therefrom, may possibly be put.

Consider, for example, the "true value" of the length of a particular gage block. In our minds we envisage the gage block as a rectangular parallelepiped, and its *length* is, of course, the distance between its two "end" faces. But it is practically certain that the particular gage block in question is not an exact rectangular parallelepiped; and that

its two end faces are not planes, nor even absolutely smooth surfaces. Shall we define the "true length" of this gage block to be the distance between the "tops" of the highest "mountains" at each end, i.e., the distance between the two "outermost points" at each end? If so, is this distance to be measured diagonally, if necessary, or parallel to the "length-wise axis" of the gage block? If the latter, then we have the problem of how this "length-wise axis" is to be defined, especially in the case of a thin gage block whose *length* corresponds to what would ordinarily be considered to be its thickness. Or shall we be, perhaps, more sophisticated, and envisage a "mean plane" at each end, which in general will not be parallel to each other, and define the length of this gage block to be the distance between two particular points on these planes. If we choose the "outermost points" we again have the problem of the direction in which the distance is to be measured. Alternatively, we might define the length of this gage block to be the distance between two strictly parallel and conceptually perfect optical flats "just touching" the gage block at each end. If so, then is the "true distance" between these flats defined in terms of wavelengths of light via the techniques of optical interferometry the "true length" of the gage block appropriate to the purposes for which the gage block is to be used, namely, to calibrate gages and to determine the lengths of other objects by *mechanical* comparisons? Furthermore, it is clear, that the intrinsic difficulty of defining the "true value" of the *length* of a particular gage block is not eliminated if, instead, we undertake to define the "true value" of the *difference in length* of two particular gage blocks, one of which is a standard, the *accepted value* of whose length is, say, m microinches *exactly*, by industry, national or international agreement.

Similar difficulties arise, of course, in the definition of the "true value" of the *mass* of a mass standard, one of which has been resolved by international agreement. In defining the "true value" of the *mass* of a particular metallic mass standard, shall *the* mass of this particular standard be envisaged as the mass of its metallic substance alone, relative to the International Prototype Kilogram, or as the mass of its metallic substance plus the mass of the air and water vapor adsorbed upon its surface under standard conditions? The difference amounts to about 45 μg in the case of a platinum-iridium standard kilogram, and becomes critical in the case of 500 mg standards. *The mass* of a mass standard is, therefore, specified in measurement science to be the mass of the metallic substance of the standard *plus* the mass of the average volume of air adsorbed upon its surface under standard conditions. Definition of the "true value" of the *mass* of a mass standard, and *a fortiori*, of the *difference in mass* of two mass standards is, therefore, a very complex matter.

W. Edwards Deming uses the expression "preferred procedure" for what we have termed an "exemplar method," and very sagely remarks that "a preferred procedure is distinguished by the fact that it supposedly gives or would give results nearest to what are needed for a particular end; and also by the fact that it is more expensive or more time consuming, or even impossible to carry out," adding that "as a preferred procedure is always subject to modification or obsolescence, we are forced to conclude that *neither the accuracy nor the bias of any procedure can ever be known in a logical sense.*" [Deming 1950, pp. 15-17.]

It should be evident from the foregoing that the "true value" of the magnitude of some property of a thing or system cannot be defined with complete absolute exactitude.

As Cassius J. Keyser has remarked, "Absolute certainty is a privilege of uneducated minds—and fanatics. It is, for scientific folk, an unattainable ideal." [Keyser 1922, p. 120.] The degree of refinement to which one will, or ought, to go in a particular instance will depend on the uses for which knowledge of the magnitude of the property concerned is needed. The "true value" of the length of a piece of cloth in everyday commerce is certainly a fuzzy concept. "Certainly we are not going to specify that the cloth shall be measured while suspended horizontally under a tension of x pounds, at an ambient temperature of y degrees and a relative humidity of z percent" [Simon 1946, p. 654]. On the other hand, a moderate degree of refinement is necessary in defining the "true length" and "true width" of the recessed area in a window sash to which a pane of glass is to be fitted. Considerably greater refinement is needed in the definition of the "true value" of the *length* of a gage block, of the *mass* of a mass standard or of the *frequency* of a frequency standard—and in the last mentioned case there is not today, I understand, complete agreement among experts on the matter.

Indeed, as is evident from the foregoing, the "true value" of the magnitude of a particular quantity is intimately linked to the purposes for which a value of the magnitude of this quantity is needed, and its "true value" cannot, in the final analysis, be defined meaningfully and usefully in isolation from these needs. Therefore, as this fact becomes more widely recognized in science and engineering, I hope that the traditional term "true value" will be discarded in measurement theory and practice, and replaced by some more appropriate term such as "target value" [6] that conveys the idea of being the value that one would like to obtain for the purpose in hand, without any implication that it is some sort of permanent constant preexisting and transcending any use that we may have for it. I have retained the traditional expression "true value" in the sequel because of its greater familiarity, but shall always mean by it the relevant "target value."

[6] "We admit the existence of systematic error—of a difference between the quantity measured (the measured quantity) and the quantity of interest (the target quantity). We ask the observations about the measured quantity. We ask our subject matter knowledge, intuition, and general information about the relation between the measured quantity and the target quantity." [Cochran, et al, 1954, p. 33.]

". . . . . Some people prefer the term 'true value', although others excoriate it as philosophically unsound.

"We could also call the reference level a 'target value'. In a way this is a bad term because it implies that it is something we want to find through the measurement process rather than something we ought to find because, like Mt. Everest, it is there. Unfortunately our desires can influence our notion of what is true, and we can even unconsciously bring the latter into agreement with the former; my use of the term 'target value' is not meant to imply that I think it legitimate to equate what we would like to see with what is there." [Murphy 1961, p. 265.]

d. Concepts of the Precision and Accuracy of a Measurement Process

By the *precision* of a measurement process we mean the degree of mutual agreement characteristic of independent measurements of a single quantity yielded by repeated applications of the process under specified conditions; and by its *accuracy* the degree of agreement of such measurements with the true value of the magnitude of the quantity concerned. In other words, the *accuracy* of a measurement process refers to, and is determined by the degree of conformity *to the truth* that is characteristic of independent measurements of a single quantity produced (or producible) by the repeated applications of the process under specified conditions; whereas its *precision* refers solely to, and is determined solely by the degree of conformity *to each other* characteristic of such measurements, irrespective of whether they tend to be close or far from the truth. Thus, *accuracy* has to do with *closeness to the truth; precision*, only with *closeness together*

This distinction between the meanings of the terms "accuracy" and "precision" as applied to measurement processes and measuring instruments is consistent with the etymological roots of these words. "Etymologically the term 'accurate' has a Latin origin meaning 'to take pains with' and refers to the care bestowed upon a human effort to make such effort what it *ought* to be, and 'accuracy' in common dictionary parlance implies freedom from mistakes or exact conformity to *truth*. 'Precise,' on the other hand, has its origin in a term meaning 'cutoff, brief, concise'; and 'precision' is supposed to imply the property of determinate limitations or being exactly and sharply defined." [Shewhart 1939, p. 124.] Thus one can properly speak of a national, state, or local law as being "precise," but not as being "accurate"—to what *truth* can it conform? On the other hand, if one spoke of a particular translation as being "accurate" this would imply a high degree of fidelity to the original "attained by the exercise of care." Whereas, to speak of it as being "precise," would imply merely that it is unambiguous, without indicating whether it is or is not correct.[7]

In spite of the distinct difference between the etymological meanings of the terms "accuracy" and "precision," they are treated as synonyms in many standard dictionaries; and Merriam-Webster [1942], after drawing the helpful distinctions quoted in the foregoing footnote, promptly topples the structure so carefully built by adding "scrupulous exactness" as an alternative meaning of "precise." Consequently it is not surprising that "There are probably few words as loosely used by scientists as *precision* and *accuracy*.—It is not unusual to find them used interchangeably in scientific writings." [Schrock 1950, p. 10.]

On the other hand, as Shewhart has remarked:

"Careful writers in the theory of errors, of course, have always insisted that accuracy involves in some way or other the difference between what is observed and what is true, whereas precision involves the concept of reproducibility of what is observed. Thus Laws, writing on electrical measurements, says: [8] 'Every experimenter must form his own estimate of the accuracy, or approach to the absolute truth obtained by the use of his instruments and processes of measurement. He must remember that a high precision, or agreement of the results among themselves, is no indication that the quantity under measurement has been accurately determined.' As another example we may take the following comment from a recent and authoritative treatise on chemical analysis: [9] 'The analyst should form the habit of estimating the probable accuracy of his work. It is a common mistake to confuse accuracy and precision. Accuracy is a measure of the degree of correctness. Precision is a measure of reproducibility in the hands of a given operator.' " [Shewhart 1939, pp. 124–125.]

More recently, Lundell, Hoffman, and their associates at the National Bureau of Standards have re-emphasized the importance of the distinction between "precision" and "accuracy":

"In discussions of chemical analysis, the terms precision and accuracy are often used interchangeably and therefore incorrectly, for precision is a measure of reproducibility, whereas accuracy is a measure of correctness. The analyst is vitally interested in both, for his results must be sufficiently accurate for the purpose in mind, and he cannot achieve accuracy without precision, especially since his reported result is often based on one determination and rarely on more than three determinations. The recipient of the analysis is interested in accuracy alone, and only in accuracy sufficient for his purposes." [Hillebrand et al., 1953, p. 3.]

It is most unfortunate that in everyday parlance we often speak of "accuracy and precision," because *accuracy* requires *precision*, but *precision* does not necessarily imply *accuracy*.

"It is, in fact, interesting to compare the measurement situation with that of a marksman aiming at a target. We would call him a precise marksman if, in firing a sequence of rounds, he were able to place all his shots in a rather small circle on the target. Any other rifleman unable to group his shots in such a small circle would naturally be regarded as less precise. Most people would accept this characterization whether either rifleman hits the bull's-eye or not.

"Surely all would agree that if our man hits or nearly hits the bull's-eye on all occasions, he should be called an accurate marksman. Unhappily, he may be a very precise marksman, but if his rifle is out of adjustment, perhaps the small circle of shots is centered at a point some distance from the bull's-eye. In that case we might regard him as an inaccurate marksman. Perhaps we should say that he is a potentially accurate marksman firing with a faulty rifle, but speaking categorically, we should have to say that the results were inaccurate." [Murphy 1961, p. 265.]

It follows from what has been said thus far that "if the precisions of two processes are the same but the biases are different, the process of smaller bias may be said to have higher accuracy while if the biases are both negligible, the process of higher precision may be said to have higher accuracy." Unfortunately, "in other cases such a simple comparison may be impossible." [ASTM 1961, p. 1760.]

[7] It is sometimes helpful to distinguish between "correct," "accurate," and "exact": "CORRECT, the most colorless term, implies scarcely more than freedom from fault or error, as judged by some (usually) conventional or acknowledged standard; . . . ACCURATE implies, more positively, fidelity to fact or truth attained by the exercise of care; . . . EXACT emphasizes the strictness or rigor of the agreement, which neither exceeds nor falls short of the fact, standard or truth; . . . PRECISE stresses rather sharpness of definition or delimitation . . ." [Merriam-Webster 1942 p. 203].

[8] Frank A. Laws, Electrical Measurements, p. 593 (McGraw-Hill, New York, N.Y., 1917).

[9] G. E. F. Lundell and J. I. Hoffman, Outlines of Methods of Chemical Analysis, p. 220 (John Wiley and Sons, New York, N.Y., 1938).

To fully appreciate the preceding statement—and especially the difficulty of comparing accuracies in some cases—let us consider figures 1 and 2, in which the origins of the scales correspond to the true value of τ of the quantity measured, so that the curves shown may be regarded as depicting the distributions of *errors* of the measurements yielded by a selection of different measurement processes. Consider first the three symmetrical distributions in the top half of figure 1. All three of these distributions are centered on zero, so that these measurement processes have no *bias*. It is evident that the process of highest precision, *c*, is also the process of highest accuracy; and that the process of least precision, *a*, is also the process of least accuracy. Since curve *b* in the upper half of figure 1 and curve *d* in the lower half have identical size and shape, the corresponding processes have the same *precision;* but process *b* is without bias, whereas process *d* has a positive bias of two units, so that process *b* is clearly the more *accurate*. (In particular we may note that whereas it is practically certain that process *b* will not yield a measurement deviating from the truth by more than two units, exactly one-half of the measurements yielded by process *d* will deviate from the truth by this much or more.) Similar remarks clearly apply to processes *c* and *e* corresponding to curve *c* in the upper half and curve *e* in the lower half of figure 1, but in this instance the superiority of process *c* relative to process *e* with respect to *accuracy* is even more marked. (In particular, we may note that whereas it is practically certain that no measurement yielded by process *c* will deviate from the truth by as much as one unit, it is practically certain that every measurement yielded by process *e* will deviate from the truth by more than one unit.)

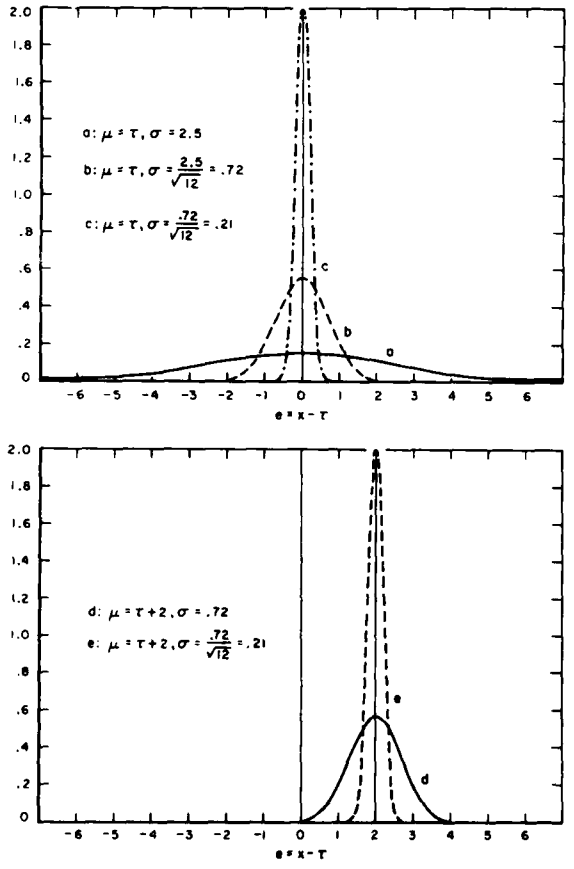


FIGURE 1. *Distributions of errors of some biased and unbiased measurement processes of various precisions.*

Figure 2, which is essentially the same as one given by General Simon [1946, fig. 1], portrays three measurement processes *A*, *B*, and *C*, differing from each other with respect to both precision and bias. Comparison of these three processes with respect to *accuracy* is not quite so simple. First, it is evident that, although process *A* has greater precision than process *B*, process *B* is the more accurate of the two. (In particular, it is practically certain that none of the measurements yielded by process *B* will deviate from the truth by more than 4 units, whereas 50 percent of the measurements from process *A* will deviate from the truth by four units or more.) Next, is process *B* more (or less) *accurate* than process *C* which is *unbiased*, but has a very low *precision*? Process *B* has a positive *bias* of two units, but has sufficiently greater *precision* than process *C* to also have greater *accuracy* than process *C*. (While approximately 50 percent of the measurements

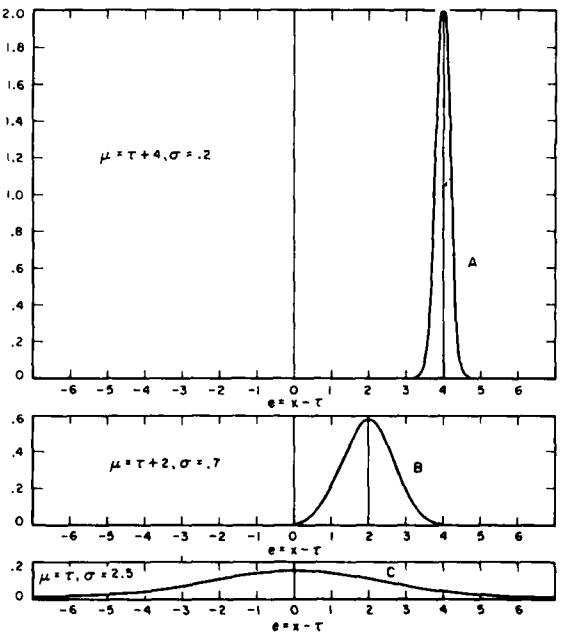


FIGURE 2. *Three measurement processes differing from each other with respect to both precision and accuracy.*

yielded by process C will deviate from the truth by more than two units (in either direction), and exactly 50 percent of the measurements yielded by process B will deviate from the truth by two units or more (in the positive direction only), it cannot be ignored that about 10 percent of the measurements yielded by process C will deviate from the truth by four units or more whereas it is practically certain that no measurement yielded by process B will deviate from the truth by as much as four units.) Similarly, it may be argued that process A, in spite of its bias, has greater *accuracy* than process C "since the range in measurements of C more than covers the corresponding ranges of A or B." [Simon 1946, p. 654.] While this conclusion that of the three measurement processes depicted in figures 2, process C has *the least accuracy, may not be entirely* acceptable to some persons, it is consistent with Gauss' dictum, in a letter to F. W. Bessel, to the effect that maximizing the probability of a zero error is less important than minimizing the "average" injurious effects of errors in general. [C. F. Gauss, 1839, pp. 146–147.]

Before leaving figure 2, we must not fail to join General Simon in remarking that "the average of a large number of measurements from [process] C will be more accurate than a similar average from either A or B" [Simon 1946, p. 654]. This point is actually illustrated in our figure 1: the three curves in the top half of figure 1 portray the distributions of errors of *single* measurements (curve *a*) of *averages of 12* measurements (curve *b*) and *averages of 144* measurements (curve *c*) from process C; and curves *d* and *e* in the lower half show the distributions of errors of *indiridual* measurements (curve *d*), and of *averages of 12* measurements (curve *e*) from process B, respectively. It is evident that *averages of 12* measurements from process C (curve *b* in upper portion of fig. 1) have not only greater *accuracy* than *individual* measurements from process B (curve *d* in lower portion of the figure), but also greater *accuracy* than *averages of 12* measurements from process B (curve *e* in lower portion).

On the other hand, it is obvious that, if our choice is between individual measurements from process C (curve *a*) and *averages of 12* measurements from process B (curve *e*), the latter will clearly provide greater *accuracy*. In brief, *a procedure with a small bias and a high precision can be more accurate than an unbiased procedure of low precision.* It is important to realize this, for in practical life it is often far better to always be quite close to the true value than to deviate all over the place in individual cases but strictly correct "on the average," like the duck hunter who put one swarm of shot ahead of the duck, and one swarm behind, lost his quarry, but had the dubious satisfaction of knowing that in theory he had hit it "on the average." This we must remember: in practical life we rarely make a very large number of measurements of a given type—we can't wait to be right on the average—our measurements must stand up in individual cases as often as possible.

Despite the foregoing, freedom from bias, that is, freedom from "large" bias, is a desirable characteristic of a measurement process. After all we want our measurements to yield us a determination that we can use as a substitute for the unknown value of a particular magnitude whose value we need for some purpose—we don't want a determination of the value of some other magnitude whose relation to the one we need is indefinitely known.

In view of the difficulty of comparing with respect to *accuracy* measurement processes that differ both in *bias* and *precision*, some writers have elected to take the easy way out by defining "*accuracy*" to be equivalent to absence of bias, saying that of two measurement processes having different biases, the process of smaller bias is the more "accurate" regardless of the relation of their respective precisions. (See, for example, Beers [1953, p. 4], Ostle [1954, p. 4], and Schenck [1961, p. 4, p. 14].) While the adoption of this concept of "accuracy" certainly makes the discussion of "accuracy" and "precision" simpler for the authors concerned, this practice is contrary to the principle of "conservation of linguistic resources," as R. B. Murphy puts it, adding: "It seems to me that the terms 'bias' and 'systematic error' are adequate to cover the situation with which they are concerned. If, nevertheless, we add the term 'accuracy' to apply again in this restricted sense, we are left wordless—at the moment at least—when it comes to the idea of over-all error. From the point of view of the need for a term it is hard to defend the view that accuracy should concern itself solely with bias. . . . [and] there is overwhelming evidence that we need a term at least for the concept of over-all error." [Murphy 1961, pp. 265–266.]

3.4. Mathematical Specification of the Precision of a Measurement Process

a. Simple Statistical Control

Let us now consider the mathematical definition of the *precision* of a measurement process under a fixed set of circumstances. By definition, the precision of a measurement process has to do with the "closeness together" that is typical of successive measurements of a single quantity generated by applications of the process under these fixed conditions. Otherwise expressed, it has to do with the typical "closeness together" of the two individual measurements constituting an arbitrary pair. If the expression "*typical* 'closeness together'" is to be meaningful, the measurements generated by repeated application of the process to the measurement of a single quantity must be homogeneous in some sense. Therefore, for the moment, let us assume that the measurement process is in a state of *simple statistical control*, so that the successive measuremets in each of the sequences (1), ($i=1, 2, 3, \ldots$), generated by the process may *all* be regarded as "observed" values of independent identically distributed random variables.

Just as we may regard each individual measurement x_{ij} in a particular sequence (1) as striving to express the value of the limiting mean μ, so also we may regard each individual difference $x_{ij}-x_{ik}$, $j \neq k$, as striving to express the characteristic spread between an arbitrary pair of measurements, x' and

x'', say. For this purpose the signs of these differences are clearly irrelevant. Therefore, by analogy with our use of a sequence of cumulative arithmetic means, (2), to achieve a mathematical formulation of the concept of a limiting mean associated with measurement of a given quantity by a particular measurement process, let us adopt the sequence of cumulative arithmetic means of the *squares* of the $n(n-1)/2$ distinct differences among the first n measurements of a particular sequence (1), for example, the sequence

$$\overline{(d^2)}_{1n} \equiv \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} (x_{1j}-x_{1k})^2, \qquad (n=2,3\ldots), \quad (3)$$

as the basis of a mathematical formulation of the concept of the precision of a measurement process.

The necessary and sufficient condition for almost sure convergence of the sequence (3) to a finite limit, say Δ^2, is that the Strong Law of Large Numbers be applicable to the sequence.

$$x_{11}^2, x_{12}^2, \ldots, x_{1j}^2 \ldots, \qquad (4)$$

consisting of the squares of the corresponding terms of the original sequence (1). (Boundedness of the x's in addition to statistical control is, for example, sufficient to ensure that the sequence (4) will also obey the Strong Law of Large Numbers.) If the Strong Law of Large Numbers is applicable to the sequence of squares (4), and if the measurement process is in a state of simple statistical control, then the cumulative arithmetic means of the squares of the measurements, that is, the sequence

$$\overline{(x^2)}_{1n} \equiv \sum_{j=1}^{n} x_{1j}^2/n, \qquad (n=1,2,\ldots), \qquad (5)$$

will almost surely tend to a limit, say S, the magnitude of which will depend on the quantity measured, the measurement process involved, but not on the "occasion" (identified by the subscript "i"). By virtue of an algebraic identity that is well known to students of mathematical inequalities, namely,

$$n\sum_{j=1}^{n} a_j^2 - \left(\sum_{j=1}^{n} a_j\right)^2 = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n}(a_j-a_k)^2, \qquad (n\geq 2) \quad (6)$$

and of the fact that the right-hand side of (6) is always positive except when the a's are all equal, it is easily seen, on dividing both sides of (6) by n^2, that S will always exceed μ^2, the square of the (almost sure) limit of the sequence (2), so that we may write $S=\mu^2+\sigma^2$, with $\sigma^2>0$. Furthermore, applying the algebraic identity (6) in reverse to the right-hand side of (3) yields the following relationship between the corresponding terms of sequences (3), (5), and (1):

$$\overline{(d^2)}_{1n} = 2\left(\frac{n}{n-1}\right)\left\{\overline{(x^2)}_{1n} - (\bar{x}_{1n})^2\right\} > 0, \qquad (n\geq 2). \qquad (7)$$

Hence, if a measurement process is in a state of simple statistical control and the Strong Law of Large Numbers is applicable to a *sequence* of squared measurements (4), then the sequence $\overline{(d^2)}_{1n}$, defined by (3), will, in view of (7), tend almost surely to a finite limit $\Delta^2=2\sigma^2$. Thus we see that σ^2, termed the *variance* of the measurement process, is the mean value of one-half of the squared difference between two arbitrary measurements x' and x'', that is,

$$\sigma^2 = \tfrac{1}{2}\overline{(x'-x'')^2}, \qquad (8)$$

and provides an indication of the *im*precision of the process. The square root of the variance, σ, is termed the *standard deviation* of the process.

It is natural, therefore, on the basis of a single sequence of n measurements of a single quantity, to take

$$s^2 \equiv \frac{1}{2}\overline{(d^2)} = \frac{1}{n(n-1)}\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}(x_j-x_k)^2 = \frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{n-1} \qquad (9)$$

as the sample estimate of the underlying variance σ^2; and the square root, s, as the sample estimate of σ.[10]

From (9), since $\bar{x}\equiv\bar{x}_n$ tends (almost surely) to μ it is evident that σ^2 is also the mean value of the squared deviations of individual measurements from the limiting mean μ of the process, that is $\sigma^2=\overline{(x-\mu)^2}$, so that the standard deviation σ may be regarded, in the language of mechanics, as the radius of gyration of the distribution of all possible measurements x about μ, the limiting mean of the process.

Remark: Mathematically the foregoing discussion can be carried out equally well in terms of the absolute (unsigned) values of the differences instead of in terms of their squares. Such an approach is, mathematically speaking, somewhat more general in that it requires for its validity merely that the Strong Law of Large Numbers be applicable to the sequence $|x_{11}|, |x_{12}|, \ldots, |x_{1j}|, \ldots$ of *absolute values* of the x_{1j} rather than to the sequence (4) of their squares. From the practical viewpoint, however, this greater generality is entirely illusory, and the mathematics of absolute values of variables is always more cumbersome than the mathematics of their squares. For example, the arithmetic mean of the absolute values of the $n(n-1)/2$ *distinct* differences among n measurements, i.e.,

$$\overline{|d|}_n \equiv \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}|x_j-x_k| \qquad (10)$$

[10] From the algebraic identity (6), it is evident that the practice in some circles of dividing $\sum_{j=1}^{n}(x-\bar{x})^2$ by n, instead of $n-1$, amounts to including each of the *distinct* squared differences $(x_j-x_k)^2$, $j\neq k$, *twice* in the summation, together with n identically zero terms $(x_j-x_k)^2$, $j=k$, each included once, and then dividing by n^2, the total number of terms (real and phantom) involved. Viewed in this light it would seem that division by $n-1$ is more reasonable, in that the inclusion of identically zero terms in the formulation of a measure of *variation* is a bit unreasonable.

is *not* expressible as a multiple of the sum of the absolute deviations of the measurements from their mean, $\sum|x_i-\bar{x}|$, and for large values of n the evaluation of (10) presents computational difficulties. The approach in terms of the absolute values of the differences also has the disadvantage from the practical viewpoint that, as we shall see in a moment, *components of imprecision* are additive in terms of squared quantities such as σ^2, so that in this sense the *variance* σ^2 is a more appropriate measure of the dispersion of the x's about their limiting mean μ than is σ itself.

Ordinarily, the magnitude of σ^2 (and, hence, of σ), unlike that of μ, depends only on the measurement process concerned and the circumstances under which it is applied, and not also on the magnitude of the quantity measured—otherwise we could not speak of a measurement process having a variance, or a standard deviation.

Since the *precision* of the process obviously decreases as the value of σ (or, of σ^2) increases, and vice versa, it is necessary to take some inverse function of σ as a measure of the precision of process. To conform with traditional usage it is necessary to regard the precision of a measurement process as *inversely proportional* to its *standard deviation* σ which is, therefore, a measure of the *imprecision* of the process. Thus, Gauss, writing in 1809, remarked that his constant $h=1/\sigma\sqrt{2}$ could properly be considered to be a measure of the precision of the observations because if, for example $h'=2h$, that is, if $\sigma'=\frac{1}{2}\sigma$, then "a double error can be committed in the former system with the same facility as a single error in the latter, in which case, according to the common way of speaking, a double degree of precision is attributed to the latter observations." [11]

The fact of the matter is, however, that:

". . . different fields have particularly favorite ways of expressing precision. Most of these measures are multiples of the standard deviation; it is not always clear which multiple is meant. . . .

"Some consider it unfortunate that precision should be stated as a multiple of standard deviation, since precision should increase as standard deviation decreases. Indeed, it would be more exact to say that standard deviation is a measure of imprecision. However, sensitivity, as we have previously indicated, suffers from this logical inversion without hurt. Perhaps we can best avoid this by saying that standard deviation is an index of precision. The habit of saying 'The precision is . . . ' is deeply rooted, and there would be understandable impatience with the notion that standard deviation should be numerically inverted before being quoted in a statement of precision." [Murphy 1961, pp. 266–267.]

In consequence the ASTM has, at least tentatively, taken the following position:

"The numerical value of any commonly used index of precision will be smaller the more closely bunched are the individual measurements of a process. As more causes are added to the system, the greater the numerical value of the index of precision will ordinarily become. If the same index of precision is used on two different processes based on the same method or intended to measure the same physical property, the process that has the smaller value of the index of precision is said to have higher precision. Thus, although the more usual indexes of precision are really direct measures of *imprecision*, this inversion of reference has been firmly established by custom. The value of the selected index of precision of a process is referred to simply as its precision or its stated precision." [ASTM 1961, p. 1759.]

As we have remarked previously, in practical work the end result of measuring some quantity or calibrating an instrument for a standard rarely consists of a single measurement of the quantity of interest. More often it is some kind of average or adjusted value, for example, the arithmetic mean of a number of independent measurements of the quantity of interest. Let us, therefore, consider the statistical properties of a sequence of *arithmetic means* of successive nonoverlapping groups of n measurements each from a sequence (1) of *individual* measurements yielded by a measurement process on a particular occasion. In other words, let us consider the sequence

$$\bar{x}_{i1}, \bar{x}_{i2}, \ldots, \bar{x}_{im}, \ldots \tag{11}$$

of distinct arithmetic means of n measurements each

$$\bar{x}_{im}=\frac{1}{n}\sum_{j=(m-1)n+1}^{mn} x_{ij}, \quad (m=1, 2, \ldots), \tag{12}$$

derived from a sequence (1) of individual measurements of a single quantity produced, or at least conceptually producible, by the measurement process concerned on, say, the ith occasion. If the "underlying measurement process" giving rise to the individual measurements x_{ij} is in a state of simple statistical control, then the "extended measurement process" giving rise to the averages $\bar{x}_{im}$ will also be in a state of simple statistical control. Consequently, the mathematical analysis of section 3.2, but with the averages $\bar{x}_{im}$ in place of the individual measurements x_{ij}, will carry through without other change. Let $\mu_{\bar{x}}$ denote the limiting mean thus associated with the "extended measurement process" giving rise to the averages $\bar{x}_{im}$ as its "individual" measurements. Since the cumulative arithmetic mean of the first m terms of the sequence (11) is the same as the cumulative arithmetic mean of the first mn terms of the sequence (1) of individual measurements, it is clear that the limiting mean $\mu_{\bar{x}}$ associated with the sequence of averages (11) is the same as the limiting mean associated with the original sequence (1) of individual measurements, that is,

$$\mu_{\bar{x}}=\mu_x=\mu. \tag{13}$$

Similarly, the mathematical analysis at the beginning of the present section, but with the individual measurements x_{ij} in (3) thru (9), replaced by the averages $\bar{x}_{im}$, carries through essentially as before. Let $\sigma_{\bar{x}}^2$ denote the variance thus associated with the "extended measurement process" giving rise to the sequence of averages (11). As in the case of the variance σ^2 of individual measurements,

[11] "Ceterum constans h tamquam mensura praecisionis observationum considerari poterit. . . . Quodsi igitur e.g., $h'=2h$, aeque facile in systemate priori error duplex committi poterit, ac simplex in posteriori, in quo casu observationibus posterioribus secundum vulgarem loquendi morem praecisio duplex tribuitur." [Gauss 1809, Art. 178; 1871, p. 233; English translation, 1857, pp. 259–260.]

so also may σ_x^2 be interpreted as the overall mean value of the squared deviation of "individual" averages $\bar{x}$ from the limiting mean μ_x of the "extended process," that is,

$$\sigma_{\bar{x}}^2=\overline{(\bar{x}-\mu_{\bar{x}})^2}=\overline{(\bar{x}-\mu)^2} \tag{14}$$

By virtue of the algebraic identity

$$(\bar{x}-\mu)^2=\left[\frac{1}{n}\sum_{j=1}^{n}x_j-\mu\right]^2=\left[\frac{1}{n}\sum_{j=1}^{n}(x_j-\mu)\right]^2$$
$$=\frac{1}{n^2}\left[\sum_{j=1}^{n}(x_j-\mu)^2+2\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}(x_j-\mu)(x_k-\mu)\right] \tag{15}$$

it is readily seen that

$$\sigma_{\bar{x}}^2=\frac{\sigma_x^2}{n}=\frac{\sigma^2}{n}. \tag{16}$$

(The mean value of a sum is always the sum of the mean values of its individual terms, so that the overall mean value of the first summation inside the brackets in the last line of (15) is simply $n\sigma_x^2$. Furthermore, in the case of independent identically distributed measurements, the overall mean value of the term involving the double summation is 0.)

Since, from (16), $\sigma_{\bar{x}}=\sigma/\sqrt{n}$, it is seen that the precision of the arithmetic mean of n *independent* measurements is proportional to $\sqrt{n}$. Hence the arithmetic mean of *4* independent measurements has *double* the precision of a single measurement; the mean of *9* independent measurements, *thrice* the precision of a single measurement; and *144* independent measurements will be required if their arithmetic mean is to have a *12-fold* increase in precision over a single measurement. (But to ask for a 12-fold increase in precision is to ask for a very considerable improvement indeed, as can be seen from a comparison of curves *a* and *c* in the top half of fig. 1.)

To serve as a reminder of the distinction between the standard deviation of an individual measurement and the standard deviation of a mean $\bar{x}$, it is customary to refer to σ as the "standard *deviation*" of a single measurement x, and to $\sigma_{\bar{x}}$ as the "standard *error*" of the (arithmetic) mean $\bar{x}$.

b. Within-Occasions Control

In the foregoing it has been assumed that the individual measurements comprising the sequences (1) corresponding to the respective "occasions," ($i=1, 2, \ldots$), could *all* be regarded as "observed values" of independent identically distributed random variables, that is, that the measurement process concerned was in a state of simple statistical control. When such is the case then any subset of n measurements is strictly comparable to any other subset of n measurements, and any two such subsets can be combined and regarded validly as a single set of $2n$ measurements. Unfortunately, as Student's comment quoted on page 167 above clearly implies, such complete homogeneity of measurement is rarely if ever met in practice. More often the situation is as described by Sir George Biddell Airy, British Astronomer Royal 1835–1881, in (to my knowledge) the first elementary book on the theory of errors and combination of observations in the English language [Airy 1861, p. 92]:

> "When successive series of observations are made, day after day, of the same measurable quantity, which is either invariable . . . or admits of being reduced by calculation to an invariable quantity . . .; and when every known instrumental correction has been applied . . .; still it will sometimes be found that the result obtained on one day differs from the result obtained on another day by a larger quantity than could have been anticipated. The idea then presents itself, that possibly there has been on some one day, or on every day, some cause, special to the day, which has produced a *Constant Error* in the measures of that day."

Sir George, however, cautions against jumping to conclusions on the basis of only a few observations:

> "The existence of a daily constant error . . . ought not to be lightly assumed. When observations are made on only two or three days, and the number of observations on each day is not extremely great, the mere fact, of accordance on each day and discordance from day to day, is not sufficient to prove a constant error. [And we should interject here that under such circumstances apparent over-all accordance is not sufficient to prove the absence of daily constant errors either.] The existence of an accordance analogous to a 'round of luck' in ordinary changes is sufficiently probable. . . . More extensive experience, however, may give greater confidence to the assumption of constant errors . . . first, it ought, in general to be established that there is possibility of error, constant on one day but varying from day to day. . . ." [Airy 1861, p. 93.]

The most useful statistical tools for this purpose are the control-chart techniques of the industrial quality control engineer. If in such a situation, a series of measurements obtained by measurement of a single quantity a number of times on each of several different days or "occasions" by a particular measurement process is plotted in the form of a *control chart for individuals* [ASTM 1951, pp. 76–78, and pp. 101, 105], the individual measurements so plotted will be seen to consist of "sections" identifiable with the subsequences (1) corresponding to the respective "occasions," ($i=1, 2, 3, \ldots$), with the measurements within sections pair-wise closer together on the average than two measurements one of which comes from one section and the other from another. Such a series of measurements is clearly "out of control." If now parallel $\bar{x}$- and R-charts are constructed from these data, based on a series of samples of equal size from *within* the respective "occasions" or "sections" *only*, i.e., excluding means $\bar{x}$ and ranges R of any samples that "straddle" two occasions, and the points on the resulting $\bar{x}$-chart are clearly "out of control," then we may infer the existence of day-by-day components of error, constant, perhaps, on one day, but varying from day to day.

If points on the R-chart constructed as described are "out of control" also, then the measurement operation concerned is in a completely unstable condition and cannot be described validly as a "measure-

ment process" at all. On the other hand, if the $\bar{x}$-chart is "out of control," but the R-chart is "in control," then we may regard the measurement process as being in a state of *within-occasions control*. ("It is usually not safe to conclude that a state of control exists unless the plotted points for at least 25 successive subgroups fall within the 3-sigma control limits. In addition, if not more than 1 out of 35 successive points, or not more than 2 out of 100, fall outside the 3-sigma control limits, a state of control may ordinarily be assumed to exist." [ASA 1958c, p. 18.]) In such a situation we postulate the existence of (at least, conceptually) different limiting means μ_i for the respective "occasions" ($i=1, 2, \ldots$), and a common *within-occasions variance* σ_w^2.

An unbiased estimate of the *within-occasions standard deviation* σ_w can be obtained, if desired, from the average range $\bar{R}$ used in constructing the R-chart, by means of the formula

$$\text{unbiased estimate of } \sigma_w = \bar{R}/d_2 \qquad (17)$$

where d_2 is the factor given in the d_2 column of table B2 of [ASTM 1951, p. 115] corresponding to the sample or subgroup size n used in constructing the R-chart.

Alternatively, if desired, an unbiased estimate of σ_w^2 can be obtained directly from the measurements involved by means of the formula

$$\text{unbiased estimate of } \sigma_w^2 = s_w^2 = \frac{\sum_{h=1}^{k}\sum_{j=1}^{n}(x_{hj}-\bar{x}_h)^2}{k(n-1)}, \qquad (18)$$

where x_{hj} denotes the jth measurement and $\bar{x}_h$ the arithmetic mean of the n measurements of the hth subgroup, respectively, and k is the number of subgroups involved in constructing the R-chart.

c. Complex or Multistage Control

When a measurement process is not in a state of simple statistical control that satisfies the criteria of within-occasions control, that is, when the $\bar{x}$-chart (and control chart for individuals) are clearly "out of control," but the 25 or more subgroup ranges plotted on the R-chart exhibit control, then it is usually of importance to ascertain whether the measurement process concerned is possibly in a state of *complex or multistage statistical control*. For this purpose four or more measurements from each of at least 25 different occasions will be needed. Taking one sample of n successive measurements, ($4 \leq n \leq 10$), from the available measurements corresponding to each of, say, $k(\geq 25)$ different "occasions," evaluate the arithmetic means $\bar{x}_i$ of these samples, ($i=1, 2, \ldots, k$), and *treating these averages as INDIVIDUAL measurements* construct a control chart for these "individuals" and parallel $\bar{x}$- and R-charts as described in [ASTM 1951, Example 22, p. 101]. If the points plotted on these three control charts exhibit control, then we "act for the present as if" the measurement process concerned is in a state of *complex or multistage statistical control* and regard the limiting means μ_i for the respective "occasions," ($i=1, 2, \ldots$) as being in a state of simple statistical control with a limiting mean μ and variance σ_b^2, termed the *between-occasions* component of variance.

If in such a situation we were to form cumulative arithmetic means such as (3) of the squares of all distinct differences between arbitrary pairs of measurements from *within* each of the respective "occasions," then such cumulative arithmetic means of squares of differences would almost surely tend to $2\sigma_w^2$ in the limit as the number of pairs included tends to infinity, where σ_w^2 is the "within-occasions variance" mentioned above in connection with "within-occasions control." If, on the other hand we were to form similar cumulative arithmetic means of the squares of differences between arbitrary pairs consisting *in each instance of one measurement from* each of two different sections, then such a cumulative arithmetic mean of squared differences would tend almost certainly to $2(\sigma_w^2+\sigma_b^2)$ as the number of "occasions" sampled tends to infinity, where σ_b^2 is the above mentioned "between-occasions variance," i.e., the variance of the limiting means μ_i for the respective "occasions" about their limiting mean μ.

If in utilizing measurements from a measurement process that is in such a state of complex statistical control, one forms an average $\bar{x}_N$ that is the arithmetic mean of a total of $N=kn$ measurements, composed of n measurements from each of k different "occasions," then the variance of $\bar{x}_N$ will be

$$\sigma_{\bar{x}_N}^2 \equiv \overline{(\bar{x}_N-\mu)^2} = \frac{1}{k}\left(\sigma_b^2+\frac{\sigma_w^2}{n}\right) \qquad (19)$$

From (19) it is clear that, if σ_b^2 is at all sizable compared to σ_w^2, then, for fixed $N=kn$, $\bar{x}_N$ will have greater precision *as a determination of* μ when based on a large number k of different occasions, with only a small number n of measurements from each occasion. Finally, setting $k=1$, we see that the mean $\bar{x}_i$, of n measurements all taken on the same occasion *considered as a determination of the overall limiting mean* μ has an *overall variance* $\sigma_x^2=\sigma_b^2+(\sigma_w^2/n)$; but *considered as a determination of* μ_i, the limiting mean for the ith occasion, its variance is only σ_w^2/n. In other words, the "standard error" of a mean such as $\bar{x}_i$ is not unique, but depends on the purpose for which it is to be used.

An unbiased estimate of the overall standard deviation $\sigma_{\bar{x}_i}$ of the arithmetic mean of n measurements taken on a single "occasion" may be obtained by the procedure of formula (17) above, if desired, using the average range $\bar{R}$ employed in constructing the R-chart corresponding to the groups of averages $\bar{x}_{in}$.

Alternatively, an unbiased estimate of the overall variance $\sigma_{\bar{x}}^2$ can be obtained directly from the means $\bar{x}_i$ used in constructing the $\bar{x}$-chart, by using the formula

$$s_{\bar{x}}^2 = \frac{\sum_{i=1}^{k} (\bar{x}_i - \bar{\bar{x}})^2}{k-1} \qquad (20)$$

where $\bar{x}_i$ is the arithmetic mean of the n successive observations from the ith "occasion," $(i=1, 2, \ldots, k)$ and $\bar{\bar{x}}$ is the arithmetic mean of these k means.

The foregoing concept of a state of *complex or multistage statistical control* can be extended readily to more complex truly "multistage" situations involving three or more "levels" of random variation.

Finally, it is evident from the foregoing that when a measurement process is in a state of complex or multistage statistical control, then the difference between two individual measurements (or the arithmetic means of n measurements) corresponding to two different "occasions" will include the difference $\mu_i - \mu_{i'}$ between the limiting means corresponding to the two particular occasions involved. In so far as such a comparison is regarded as a unique individual case, the difference $\mu_i - \mu_{i'}$ is a fixed constant and hence a systematic error affecting this comparison. On the other hand, if the difference between these two individual measurements (or these two arithmetic means) is regarded only as a typical instance of the outcomes that might be yielded by the same measurement process on *other* pairs of occasions, then the difference $\mu_i - \mu_{i'}$ may be regarded as a random component having a zero mean and variance $2\sigma_b^2$.

It goes without saying, of course, that if a control-chart analysis of the type described above is undertaken for the purpose of ascertaining whether the process is in a state of complex control, but the points plotted on the $\bar{x}$-chart are clearly "out of control," then the measurement process concerned cannot be regarded as statistically stable from occasion to occasion, and should be used only for *comparative measurement* within-occasions. Even when such a measurement process is used solely for comparative measurement within "occasions," it needs to be shown that comparative measurements or *fixed differences* are in a state of (simple or complex) statistical control, if this measurement process is to be generally valid in any absolute sense. Thus in the case of the thermometer calibration procedure mentioned in section 2.4 above, one needs to examine the results of repeated measurement, occasion after occasion, of the difference between two standard thermometers S_1 and S_2 of proven stability in order to determine whether the process is or is not in a state of simple or complex statistical control.

3.5. Difficulty of Characterizing the Accuracy of a Measurement Process

Unfortunately, there does not exist any single comprehensive measure of the accuracy (or *in*accuracy) of a measurement process (analogous to the standard deviation as a measure of its imprecision) that is really satisfactory. This difficulty stems from the fact that "accuracy," like "true value," seems to be a reasonably definite concept on first thought, but as soon as one attempts to specify exactly what one means by "accuracy" in a particular situation, the concept becomes illusive; and in attempting to resolve the matter one comes face to face, sooner or later, with the question: "Accurate" for what *purpose*?

Gauss, in his second development (1821–1823) of the Method of Least Squares clearly recognized the difficulty of characterizing sharply the "accuracy" of any particular procedure:

> "Quippe quaestio haec per rei naturam aliquid vagi implicat, quod limitibus circumscribi nisi per principium aliquatenus arbitrarium nequit . . . neque demonstrationibus mathematicis decidenda, sed libero tantum arbitrio remittenda." [12] [Gauss 1823, Part I, Art. 6.]

Gauss himself proposed [loc. cit.] that the *mean square error* of a procedure—that is, $\sigma^2+(\mu-\tau)^2$, where σ is its *standard deviation;* and $\mu-\tau$, its *bias*—be used to characterize its accuracy. While *mean square error* is a useful criterion for comparing the relative accuracies of measurement processes differing widely in both precision and bias, it clearly does not "tell the whole story." For example, if one were to adopt the principle that measurement processes having the same mean square error were equally "accurate," then one would be obliged to consider the measurement processes corresponding to the three curves shown in figure 3 as being of equal

[12] I am grateful to my colleague Franz Alt for the following literal translation of these phrases:

"For this question implies, by the very nature of the matter, something vague which cannot be clearly delimited except by somewhat arbitrary principle . . . nor can it be decided by mathematical demonstrations, but must be left to mere arbitrary judgment."

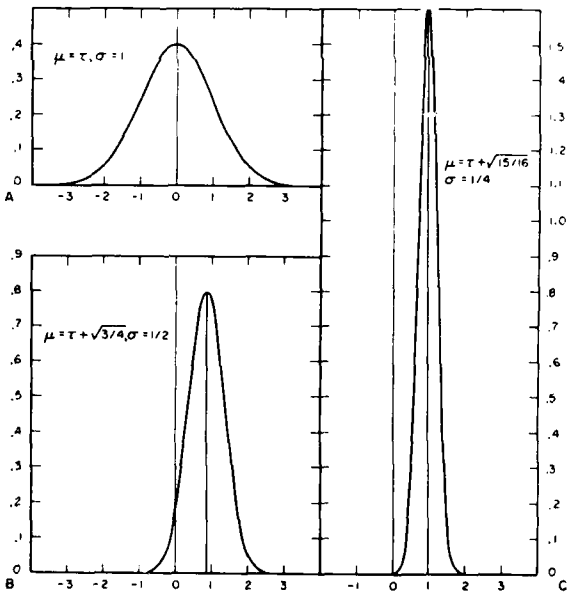


FIGURE 3. *Three distributions differing with respect to both precision and accuracy but with the same mean square error.*

accuracy, whereas for many purposes one would regard process C (portrayed to the right) as the "most accurate," in spite of the fact that the chances of scoring a "bull's eye" or "near miss" are greater in the case of process A shown in the upper left.

Alternatively, if one were to say that two measurement processes were equally accurate when exactly the same proportion P of the measurements of each lay within $\pm\delta$ units from the true value, then for $P=0.5$ one would be obliged to say that the measurement processes corresponding to curves e and d in the lower half of figure 1 were equally accurate, and that the measurement process corresponding to curve a in the upper half of the same figure was slightly more accurate than either e or d. Or, taking $P=0.95$, one would be obliged to say that the measurement processes corresponding to the three curves shown in figure 4 were equally accurate. From these, and other cases easily constructed, it is readily seen that it is unsatisfactory to regard two measurement processes as being equally accurate if the same specified fraction P of the measurements produced by each lie within the same distance from the true value.

Thus one is led by the force of necessity to the inescapable conclusion that ordinarily (at least) two numbers are needed to adequately characterize the accuracy of a measurement process. And this has been recognized by the American Society for Testing and Materials in their recent recommendations [ASTM 1961, pp. 1759–1760]:

> "Generally the index of accuracy will consist of two or more different numbers. Since the concept of accuracy embraces not only the concept of precision but also the idea of more or less consistent deviation from the reference level (systematic error or bias), it is preferable to describe accuracy by separate values indicating precision and bias."

The fact of the matter is that *two* numbers ordinarily suffice only because the "end results" of measurement and calibration programs are usually averages or adjusted values based on a number of independent 'primary measurements," and such averages and adjusted values tend to be normally distributed to a very good approximation when four or more "primary measurements" are involved. This is illustrated by figure 5, which shows the distributions of individual measurements of two unbiased measurement processes with identical standard deviations but having *uniform* and *normal* "laws of error," respectively, together with the corresponding distributions of arithmetic means of 4 independent measurements from these respective processes—these latter two distributions are depicted by a single curve because the differences between the two distributions concerned are far less than can be resolved on a chart drawn to this scale. Since both of the processes concerned are unbiased, "accuracy" thus becomes only a matter of "precision"—or does it?—both curves for $n=1$ have the same standard deviation, do they reflect equal "accuracy"? Would not the answer depend on the advantages to be gained from small errors balanced against the seriousness of large errors, in relation to the purpose for which a single measurement from one or the other is needed? But "the problem" disappears nicely if averages of 4 measurements are to be used.

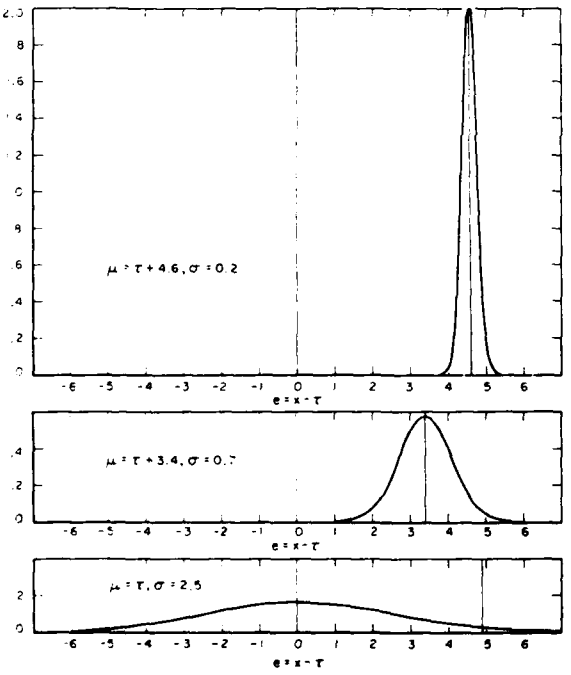


FIGURE 4. *Three measurement processes differing in bias and precision but having 95 percent of their individual measurements within ±4.9 units from the true value τ.*

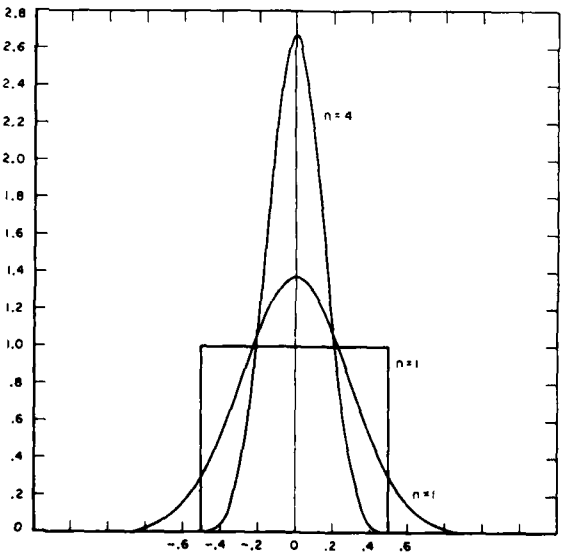


FIGURE 5. *Uniform and normal distributions of individual measurements having the same mean and standard deviation, and the corresponding distribution(s) of arithmetic means of four independent measurements.*

4. Evaluation of the Precision, and of Credible Bounds to the Systematic Error of a Measurement Process

As we have just seen, two numbers are ordinarily needed to characterize the accuracy of a measurement process, the one indicating its *precision*, and the other its *bias*. In practice, however, the bias of a measurement process is unknown and unknowable because the "true values" of quantities measured are almost always unknown and unknowable. The principle exception is when one is measuring a difference that is by hypothesis identically zero. If the bias of a measurement process could be, and were known exactly, then one would of course subtract it off as a "correction" and thus dispose of it entirely. Since ordinarily we cannot expect to know the exact magnitude of the bias of a measurement process, we are forced in practice to settle for credible bounds to its likely magnitude—much as did Steyning and the thief in chapter VI of Kipling's story, *Captains Courageous:* "Steyning tuk him for the reason that the thief tuk the hot stove—bekaze for there was nothing else that season". Consequently, neither the *bias* nor the *accuracy* of any measurement process, or method of measurement, can ever be known in a logical sense. The *precision* of a measurement process, however, can be measured and known. (Compare Deming [1950, p. 17].)

4.1. Evaluation of the Precision of a Measurement Process

In the foregoing we have stressed that a measurement operation to qualify as a *measurement process* must have attained a state of statistical control; and that until a measurement operation has been "debugged" to the extent that it has attained a state of statistical control, it cannot be regarded in any logical sense as measuring anything at all. It is also clear, from our discussion of the control-chart techniques for determining whether in any given instance one is entitled to "act for the present as if" a state of statistical control has been attained, that a fairly large amount of experience with a particular measurement process is needed before one can resolve the question in the affirmative. Once a measurement process has attained a state of statistical control, and so long as it remains in this state, then an estimate of the *standard deviation* of the process can be obtained from the data employed in establishing control, as we have indicated above.

Since the precision of a measurement process refers to, and is determined by the characteristic "closeness together" of successive independent measurements of a single magnitude generated by repeated application of the process under specified conditions, it is clearly necessary in determining whether a measurement operation is or is not in a state of statistical control, and in evaluating its precision to be reasonably definite on what variations of procedure, apparatus, environmental conditions, observers, operators, etc., are allowable in "repeated applications" of what will be considered to be the same measurement process applied to the measurement of the same quantity under the same conditions. If whatever measure of the precision and bounds to the bias of the measurement process we may adopt are to provide a realistic indication of the accuracy of this process in practice, then the "allowable variations" must be of sufficient scope to bracket the range of circumstances commonly met in practice. Scientists and engineers commonly append "probable errors" or "standard errors" to the results of their experiments and tests. These measures of imprecision are supposed to indicate the extent of the reproducibility of these experiments or tests under "essentially the same conditions," but there are great doubts whether the "probable errors" and "standard errors" generally presented actually have this meaning. The fault in most cases is not with the statistical formulas and procedures used to compute such probable errors or standard errors from the measurements in hand, but rather with the limited scope of the "conditions" sampled in taking the measurements.

a. Concept of a "Repetition" of a Measurement

As a very minimum, a "repetition" of a measurement by the same *measurement process* should "leave the door open" to, and in no way inhibit changes of the sort that would occur if, on termination of a given series of measurements, the data sheets were stolen and the experimenter were to repeat the series as closely as possible with the same apparatus and auxiliary equipment following the same instructions. In contrast, a "repetition" by the same *method of measurement* should permit and in no way inhibit the natural occurrence of such changes as will occur if the experimenter were to mail to a friend complete details of the apparatus, auxiliary equipment, and experimental procedure employed—i.e., the written text specification that defines the "method of measurement" concerned—and the friend, using apparatus and auxiliary equipment of the same kind, and following the procedural instuctions received to the best of his ability, were then, after a little practice, to attempt a repetition of the measurement of the same quantity. Such are the extremes, but there is a "gray region" between in which there is not to be found a sharp line of demarcation between the "areas" corresponding to "repetition" by the *same measurement process*, and and to "repetition" by the *same method of measurement*.

Let us consider "repetitions" by the same *measurement process* more fully. Such repetitions will undoubtedly be carried out in the same place, i.e., in the same laboratory, because if it is to be the same measurement process, the very same apparatus *must be used*. But a "repetition" cannot be carried out at the same *time*. How great a lapse of time should be allowed, nay *required*, between "repetitions"? This is a crucial question. Student gives an answer in a passage from which we quoted above [Student 1917, p. 415]:

"Perhaps I may be permitted to restate my opinion as to the best way of judging the accuracy of physical or chemical determinations.

"After considerable experience I have not encountered any determination which is not influenced by the date on which it is made; from this it follows that a number of determinations of the same thing made on the same day are likely to lie more closely together than if the repetitions had been made on different days.

"It also follows that if the probable error is calculated from a number of observations made close together in point of time, much of the secular error will be left out and for general use the probable error will be too small.

"Where then the materials are sufficiently stable it is well to run a number of determinations on the same material through any series of routine determinations which have to be made, spreading them over the whole period."

Another important question is: Are "repetitions" by the same measurement process, to be limited to repetitions by the same observers and operators, using the same auxiliary equipment (bottles of reagents, etc.); or enlarged to include repetitions with nominally equivalent auxiliary equipment, by various but equivalently trained observers and operators? I believe that everyone will agree that substitution, and certainly replacement, of bottles of reagents, of batteries as sources of electrical energy, etc., by "nominally equivalent materials" must be allowed. And any calibration laboratory having a large amount of "business" will certainly, in the long run at any rate, have to face up to allowing changes, even replacement of observers and operators—and, ultimately, even of apparatus.

A very crucial question, not always faced squarely, is: in complete "repetitions" by the same measurement process, are such "repetitions" to be limited to those intervals of time over which the apparatus is used "as is" and "undisturbed," or extended to include the additional variations that almost always manifest themselves when the apparatus is disassembled, cleaned, reassembled, and readjusted? Unless such disassembly, cleaning, reassembly, and readjustment of apparatus is permitted among the allowable variations affecting a "repetition" by the same measurement process, then there is very little hope of achieving satisfactory agreement between two or more measurement processes in the same laboratory that differ only in their identification with different pieces of apparatus of the same kind. In practice it is found that statistical control can be attained and maintained under such a broad concept of "repetition" only through the use of reference standards of proven stability. Furthermore, by thus more squarely facing the issue of the scope of variations allowable with respect to "repetitions" by the same measurement process, we shall go a long way toward narrowing the gap between a "repetition" by the same measurement process and by the same method of measurement.

As we have said before, if whatever measures of the precision and bias of a measurement process we may adopt are to provide a realistic indication of the accuracy of this process in practice, then the "allowable variations" must be of sufficient scope to bracket the range of circumstances commonly met in practice. Furthermore, any experimental program that aims to determine the precision and systematic error, and thence the accuracy of a measurement process, must be based on an appropriate random sampling of this "range of circumstances," if the usual tools of statistical analysis are to be strictly applicable. Or as Student put it, "the experiments must be capable of being considered to be a *random* sample of the population to which the conclusions are to be applied. Neglect of this rule has led to the estimate of the value of statistics which is expressed in the crescendo 'lies, damned lies, statistics'." [Student 1926, p. 711.]

When adequate random sampling of the appropriate "range of circumstances" is not feasible, or even possible, then it is necessary to compute, by extrapolation from available data, a more or less subjective estimate of the "precision" of the end results of a measurement operation, to serve as a substitute for a direct experimental measure of their "reproducibility." Youden [1962d] calls this "approach the 'paper way' of obtaining an estimate of the [precision]." Its validity, if any, "is based on subject-matter knowledge and skill, general information, and intuition—but not on statistical methodology" [Cochran et al. 1953, p. 693].

b. Some Examples of Realistic "Repetitions"

As Student remarked [1917, p. 415], "The best way of judging the accuracy of physical or chemical determination . . . [when] the materials are sufficiently stable . . . is . . . to run a number of determinations on the same material thru any series of routine determinations which have to be made, spreading them over the whole period." To this end, as well as to provide an overall check on procedure, on the stability of reference standards, and to guard against mistakes, it is common practice in many calibration procedures, to utilize two or more reference standards as part of the regular calibration procedure.

The calibration procedure for *liquid-in-glass thermometers*, referred to in section 2.4 above, is a case in point. A measurement of the difference between the two standards S_1 and S_2 is obtained as by-product of the calibration of the four test thermometers T_1, T_2, T_3, and T_4 in terms of the (corrected) readings of the two standards. It is such remeasurements of the difference between a pair of standard thermometers from "occasion" to "occasion" that constitutes realistic "repetitions" of the calibration procedure. The data yielded by these "repetitions" are of exactly the type needed (a) to ascertain whether or not the process is in a state of statistical control; and if so, (b) to determine its overall standard deviation.

Similarly, in the calibration of *laboratory standards of mass* at the National Bureau of Standards, "known standard weights are calibrated side-by-side with [the] unknown weights" [Almer et al., 1962, p. 33]. Indeed, weights whose values are otherwise determined "are not said to have been 'calibrated'. That term is reserved for measurements based on at least two mass standards." [loc. cit., p. 43.] In the specimen work sheets exhibited by Almer et al., the auxiliary standards involved are those from the Bureau's "NH series" of reference standards known

by the designations NH50, NH20, and NH10, respectively. It is the measurements obtained in routine calibrations of the differences between the values of these standards and their accepted values that not only provide valuable checks on day-to-day procedure, but also serve as the basis for determination of the overall standard deviation of this calibration process.

A third example is provided by the method followed at the National Bureau of Standards for testing *alternating-current watthour meters*, which has been described in some detail by Spinks and Zapf [1954]. Four reference watthour meters are involved. One of these, termed "the Standard Watthour Meter," is located in the device portrayed in figure 1 of the paper by Spinks and Zapf. The other three are located in a temperature-controlled cabinet. A "test" of a watthour meter sent to the Bureau involves *not only* a comparison of this watthour meter with the Standard Watthour Meter, but also comparisons of each of the Comparison Standard Watthour Meters with the Standard Watthour Meter. It is from the data yielded by these intercomparisons of the Standard Watthour Meter and the Comparision Standard Watthour Meters that the standard deviation of this test procedure is evaluated. Spinks and Zapf's section on "Precision and Accuracy Attainable" is notable for its exceptional lucidity as well as for its completeness with respect to relevant details.

Some additional examples of realistic "repetitions" are discussed by Youden [1962c].

4.2. Treatment of Inaccuracy Due to Systematic Errors of Assignable Origins but of Unknown Magnitudes

As we remarked in section 3.3b above, the systematic error of a measurement process will ordinarily have both constant and variable components. For convenience of exposition, it is customary to regard the individual components of the overall systematic error of a measurement or calibration process as elemental or constituent "systematic errors" and to refer to them simply as "systematic errors," for short. Included among such "systematic errors" affecting a particular measurement or calibration process are: ". . . all those errors which cannot be regarded as fortuitous, as partaking of the nature of chance. They are characteristic of the system involved in the work; they may arise from errors in theory or in standards, from imperfections in the apparatus or in the observer, from false assumptions, etc. To them, the statistical theory of error does not apply." [Dorsey 1944, p. 6; Dorsey and Eisenhart 1953, p. 104.]

The overall systematic error of a measurement process ordinarily consists of elemental "systematic errors" due to both assignable and unassignable causes. Those of unknown (not thought of, not yet identified, or as yet undiscovered) origin are always to be feared; allowances can be made only for those of recognized origin.

Since the "known" systematic errors affecting a measurement process ascribable to specific origins are ordinarily determinate in origin only, their individual values ordinarily being unknown both with respect to sign and magnitude, it is not possible to evaluate their algebraic sum and thereby arrive at a value for the overall systematic error of the measurement process concerned. In consequence, it is necessary to arrive at bounds for each of the individual components of systematic error that may be expected to yield nonnegligible contributions, and then from these bounds arrive at credible bounds to their combined effect on the measurement process concerned. Both of these steps are fraught with difficulties.

Determination of reasonable bounds to the systematic error likely to be contributed by a particular origin or assignable cause necessarily involves an element of judgment, and the limits cannot be set in exactitude. By assigning ridiculously wide limits, one could be practically certain that the actual error due to a particular cause would never lie outside of these limits. But such limits are not likely to be very helpful. The narrower the range between the assigned limits, the greater the uneasiness one feels that the assigned limits will not include whatever systematic error is contributed by the cause in question. But a decision has to be made; and on the basis of theory, other related measurements, a careful study of the situation in hand, especially its sensitivity to small changes in the factor concerned, and so forth, "the experimenter presently will feel justified in saying that he feels, or believes, or is of the opinion," that the systematic error due to the particular source in question does not exceed such and such limits, "meaning thereby, since he makes no claim to omniscience, that he has found no reason for believing" that it exceeds these limits. In other words, 'nothing has come to light in the course of the work to indicate" that the systematic error concerned lies outside the stated range. [Dorsey 1944, pp. 9–10; Dorsey and Eisenhart, 1953, pp. 105–107.]

This being done to each of the recognized potential sources of systematic error, the problem remains how to determine credible bounds to their combined effect. Before considering this problem in detail, it will be helpful to digress for a moment, to consider an instructive example relating to the combined effect of constant errors in an everyday situation.

a. An Instructive Example

Consider the hypothetical situation of an individual who is comparing his checkbook balance with his bank statement. To this end he needs to know the total value of his checks outstanding. Loathing addition, or perhaps, simply to save time, he adds up only the dollars, neglecting the cents, and thus arrives at a total of, say, $312, for 20 checks outstanding. Adding a correction of 50 cents per check, or $10 in all, he takes $322 as his estimate. Within what limits should he consider the error of this estimate to lie?

The round-off error cannot exceed ± 50 cents per

check, so that barring mistakes in addition, he can be absolutely certain that the total error of his estimate does not exceed ±$10. But these are extremely pessimistic limits: they correspond to every check being in error by the maximum possible amount and all in the same direction. (Actually the maximum possible positive error is 49 cents per check or +$9.80 in all.)

To be conservative, but not so pessimistic, one might "allow" a maximum error of ± 50 cents per check, but consider it reasonable to regard their signs as being equally likely to be plus or minus. In this way one would be led to conclude "with probability 0.95" that the total error lies between ± $4.00; or "with probability 0.99," between ± $6.00, as shown in the column headed "binomial" in table 1, for n = 20.

TABLE 1. *Limits of error of a sum of n items indicated by various methods of evaluation*

n	Absolute ±	Binomial*		Uniform		Triangular		Normal, 2σ=0.5		Normal, 3σ=0.5	
		0.95 ±	0.99 ±	0.95 ±	0.99 ±	0.95 ±	0.99 ±	0.95 ±	0.99 ±	0.95 ±	0.99 ±
1	0.50	0.50	0.50	0.48	0.50	0.39	0.45	0.49	0.64	0.33	0.43
2	1.00	1.00	1.00	0.78	0.90	0.56	0.71	0.69	0.91	0.46	0.61
3	1.50	1.50	1.50	0.97	1.19	0.69	0.88	0.85	1.12	0.57	0.74
4	2.00	2.00	2.00	1.12	1.41	0.80	1.03	0.98	1.29	0.65	0.86
5	2.50	2.50	2.50	1.25	1.60	0.89	1.15	1.10	1.44	0.73	0.96
6	3.00	2.00	3.00	1.38	1.76	0.98	1.29	1.20	1.58	0.80	1.06
7	3.50	2.50	3.50	1.49	1.91	1.06	1.39	1.30	1.70	0.86	1.14
8	4.00	3.00	3.00	1.59	2.05	1.13	1.49	1.39	1.82	0.92	1.21
9	4.50	2.50	3.50	1.69	2.18	1.20	1.58	1.47	1.93	0.98	1.29
10	5.00	3.00	4.00	1.78	2.31	1.27	1.66	1.55	2.04	1.03	1.36
15	7.50	3.50	4.50	2.19	2.88	1.55	2.04	1.90	2.49	1.27	1.66
20	10.00	4.00	6.00	2.53	3.33	1.79	2.35	2.19	2.88	1.46	1.92
25	12.50	4.50	6.50	2.83	3.72	2.00	2.63	2.45	3.22	1.63	2.15
30	15.00	5.00	7.00	3.10	4.07	2.19	2.88	2.68	3.53	1.79	2.35
40	20.00	6.00	8.00	3.58	4.70	2.53	3.33	3.10	4.07	2.07	2.72
50	25.00	7.00	9.00	4.00	5.26	2.83	3.72	3.46	4.55	2.31	3.04
60	30.00	8.00	10.00	4.38	5.76	3.10	4.07	3.80	4.99	2.53	3.33

*The results are not monotonic due to the discreteness of the distribution.

Alternatively, one might consider it to be more "realistic" to regard the individual errors as independently and uniformly distributed between −50 cents and +50 cents, concluding "with probability 0.95" that the total error does not exceed ±$2.53; or "with probability 0.99," is not greater then ±$3.33—as shown in the columns under the heading "uniform" in table 1. It is clear that a considerable reduction in the estimate of the total error is achieved by this approach.

Strictly speaking, the foregoing analyses via the theory of probability are both inapplicable to the problem at hand: each round-off error is a fixed number between ±50 cents, and their sum is a fixed number between ±$10. If it were true that round-off errors in such cases were uniformly distributed between ±50 cents, then, if one made a habit of evaluating limits of error according to this procedure, one could expect the limits of error so calculated to include the true total error in 95 percent, or 99 percent of the instances in which this procedure was used in the long run. Round-off errors in such cases are almost certainly not uniformly distributed between ±50 cents. (Many items are priced these days at $2.98 etc., and this will distort the distribution of the cents-portion of one's bills but added sales taxes no doubt have a "smoothing" effect.)

Nevertheless, I believe that you will agree that if, in the hypothetical case under discussion, the checkbook balance, with an allowance of $522 for checks outstanding, failed to agree with the bank statement to within $2.53 (or $3.33), our "friend" would do well to check into the matter more thoroughly. And, alternatively, if his checkbook balance so adjusted, and the bank statement, agreed to within $2.53 (or $3.33), it would be reasonably "safe" for him to "act for the present as if" his balance and the bank statement were in agreement. (See Eisenhart [1947a, p. 218] for discussion of a similar example relating to computation with logarithms.)

b. Combination of Allowances for Systematic Errors

The foregoing example suggests that a similar procedure be used for arriving at credible limits to the likely overall effect of systematic errors due to a number of different origins. A number of additional difficulties confront us, however, in this case. To begin with, in view of the inexactness with which bounds can ordinarily be placed on each of the individual components of systematic error, it is not possible to say with absolute certainty that their combined effect lies between the sum of the positive bounds and the sum of the negative bounds.

Second, even if it were possible to scale the situation so that the bounds for each of the components of systematic error was the same, say, ±Δ, there would still remain the problem of translation into an appropriate probability calculus. Most persons would, I believe, regard the "binomial" approach (corresponding to equal probability of maximum error in either direction), as too pessimistic; and the approach via a uniform distribution of error, as a bit conservative, on the grounds that one intuitively feels that the individual errors are somewhat more likely to lie near the centers than near the ends of their respective ranges. Therefore, one might attempt to simulate this "feeling" by assuming the "law of error" to be an isosceles triangle centered at zero and ends at ±Δ; or, more daringly, by assuming the "law of error" to be approximately normal with Δ corresponding to 2 "σ" or even 3 "σ."

Unfortunately whatever "probability limits" may be placed upon the combined effects of several independent systematic errors by these procedures are quite sensitive to the assumption made at this stage, as is evident from table 1. Therefore, anyone who uses one of these methods for the "combination of errors" should indicate explicitly which of these (or an alternative method) he has used. When (a) the number of systematic errors to be combined is large, (b) the respective ranges are approximately equal in size, and (c) one feels "fairly sure" that the individual errors do not fall outside of their respective ranges, then my personal feeling is that the "uniform" method is probably a wee bit conservative but "safe"; the triangular method is a bit "too daring"; the normal method with "σ"$=\Delta/3$ ordinarily "much too daring"; but the normal method with "σ" $=\Delta/2$, probably "not too daring." When (b) and (c) hold but n is small, then it will probably be safe to use the "uniform" method with "Δ" taken equal to the average of the individual ranges. Other cases, e.g., when n is large but, say, one or two of the ranges is (are) much larger than the others and tend(s) to dominate the situation, requires special consideration which is beyond the scope of the present paper.

4.3. Expression of the Inaccuracy of a Measurement Process

By whatever means credible bounds to the likely overall systematic error of the measurement process are obtained they should not be combined (by simple addition, by "quadrature," or otherwise) with an experimentally determined measure of its standard deviation to obtain an overall index of its accuracy (or, more correctly, of its *in*accuracy). Rather (a) the standard deviation of the process and (b) credible bounds to its systematic error should be stated separately, because, as we showed in figure 3, a measurement process having standard deviation $\sigma=0.25$ and a bias $\Delta=\sqrt{15/16}=0.97$ is for most purposes "more accurate" than a measurement process having zero bias and standard deviation $\sigma=1$, so that a process with $\sigma=0.25$ and a bias *less than* ± 0.97 will *a fortiori* be "more accurate."

Finally, if the uncertainties in the assigned value of a national standard or of some fundamental constant of nature (e.g., in the *volt as maintained at the National Bureau of Standards*, or in the speed of light c, or in the acceleration of gravity g on the Potsdam basis) is an important potential source of systematic error affecting the measurement process, no allowance for possible systematic error from this source should be included ordinarily in evaluating overall bounds to the systematic error of the measurement process. Since the error concerned, what ever it is, affects all results obtained by the method of measurement involved, to include an allowance for this error would be to make everybody's results appear unduly inaccurate relative to each other. Instead, in such instances one should state (a) that results obtained by the measurement process concerned are in terms of the volt (or the watthour, or the kilogram, etc.) "as maintained at the National Bureau of Standards" [McNish and Cameron 1960, p. 102], or "correspond to the speed of light $c=2.997925\times10^{10}$ cm/sec. *exactly*," say; and (b) that the indicated bounds to the systematic error of the process are exclusive of whatever errors may be present from this (or these) source(s). Given such information, experts can make such additional allowances, as may be needed, in fundamental scientific work; and comparative measurements within science and industry within the United States will not appear to be less accurate than they very likely are for the purposes for which they are to be used.

It is a pleasure to acknowledge the technical assistance of Janace A. Speckman in several phases of the preparation of this paper.

5. Bibliography

Airy, George Biddell (1861), On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations (Macmillan and Co., Cambridge and London).

Almer, H. E., L. B. Macurdy, H. S. Peiser, and E. A. Weck (1962), Weight calibration schemes for two knife-edge direct-reading balances, J. Research NBS **66C** (Eng. and Instr.) No. 1, pp. 33–44.

American Standards Association (1958a), Guide for quality control, American Standard Z 1.1–1958. (American Standards Association, 70 East Forty-fifth St., New York 17, N.Y.)

American Standards Association (1958b), Control chart method of analyzing data, American Standard Z 1.2–1958. (American Standards Association, 70 East Forty-fifth St., New York 17, N.Y.)

American Standards Association (1958c), Control chart method of controlling quality during production. American Standard Z 1.3–1958. (American Standards Association, 70 East Forty-fifth St., New York 17, N.Y.)

American Society for Testing Materials (1951), ASTM Manual on Quality Control of Materials Special Technical Publication **15-C** (American Society for Testing Materials 1916 Race St., Philadelphia 3).

American Society for Testing and Materials (1961), Use of the terms precision and accuracy as applied to measurement of a property of a material, ASTM Designation: E 177–61T. Reprinted from ASTM Standards, Pt 11, pp. 1758–1766.

Baird, D.C. (1962), Experimentation: An Introduction to Measurement Theory and Experiment Design, (Prentice-Hall, Inc., Englewood Cliffs, N.J.).

Beers, Yardley (1953), Introduction to the Theory of Error, (Addison-Wesley Publishing Co., Cambridge 42, Mass.).

Bicking, Charles A. (1952), The reliability of measured values—an illustrative example. Photogrammetric Engineering **XVIII,** pp. 554–558.

Cameron, J. M. (1951), The use of components of variance in preparing schedules for the sampling of baled wool, Biometrics **7,** pp. 83–96.

Chauvenet, William (1868), A Manual of Spherical and Practical Astronomy Vol. II, 4th edition, (J. B. Lippincott and Co., Philadelphia).

Cochran, William G., Frederick Mosteller, and John W. Tukey (1953), Statistical problems of the Kinsey report, J. Am. Stat. Assoc. **48,** pp. 673–716.

Cochran, William G., Frederick Mosteller, and John W. Tukey (1954), Principles of sampling, J. Am. Stat. Assoc. **49,** pp. 13–35.

Crow, Edwin L. (1960), An analysis of the accumulated error in a hierarchy of calibrations. IRE Trans. Instr. **I-9,** pp. 105–114.

Deming, W. Edwards and Raymond T. Birge (1937), On the Statistical Theory of Errors, reprinted from Reviews of Modern Physics **6,** pp. 119–161 (1934) with additional notes dated 1937 (The Graduate School, U.S. Department of Agriculture, Washington 25, D.C.).

Deming, W. Edwards (1943), Statistical Adjustment of Data (John Wiley & Sons, Inc., New York, N.Y.)

Deming, W. Edwards (1950), Some Theory of Sampling (John Wiley & Sons, New York, N.Y.).

Dorsey, N. Ernest (1944), The velocity of light, Transactions American Philosophical Society **XXXIV,** pp. 1–110.

Dorsey, N. Ernest and Churchill Eisenhart (1953), On absolute measurement, The Scientific Monthly **LXXVII,** pp. 103–109.

Eisenhart, Churchill (1947a), Effects of rounding or grouping data, Chapter 4 of Techniques of Statistical Analysis, edited by C. Eisenhart, M. W. Hastay, W. A. Wallis (McGraw-Hill Book Co., New York, N.Y.).

Eisenhart, Churchill (1947b), Planning and interpreting experiments for comparing two standard deviations, Chapter 8 of Techniques of Statistical Analysis, edited by C. Eisenhart, M. W. Hastay, W. A. Wallis (McGraw-Hill Book Co., New York, N.Y.).

Eisenhart, Churchill (1949), Probability center lines for standard deviation and range charts, Industrial Quality Control **vi,** pp. 24–26.

Eisenhart, Churchill (1952), The reliability of measured values—fundamental concepts, Photogrammetric Engineering **XVIII,** pp. 542–554 and 558–565.

Eisenhart, Churchill (1962), On the realistic measurement of precision and accuracy, ISA Proceedings of the Eight National Aero-Space Instrumentation Symposium held in Washington, May 1962, pp. 75–83.

Feller, William (1957), An Introduction to Probability Theory and its Applications, Vol. 1, 2d edition (John Wiley & Sons, New York, N.Y.).

Galilei, Galileo (1638), Discorsi e Dimostrazioni Matematiche Intorno a Due Nuove Scienze, Leiden.

Galilei, Galileo (1898), Discorsi e Dimostrazioni Matematiche Intorno a Due Nuove Scienze, Le Opere di Galileo Galilei (Edizione Nazionale) **VIII,** pp 39–448, Firenze.

Galilei, Galileo (1914), Dialogues Concerning Two New Sciences, translated by Henry Crew and Alfonso de Salvio, with an Introduction by Antonio Favaro (The Macmillan Co., New York, N.Y.).

Gauss, C. F. (1809), Theoria Matus Corporum Coelestium in Sectionibus Conicis Solem Ambientium, Frid. Perthes et I. H. Besser, Hamburg; reprinted in Carl Friedrich Gauss Werke, Band **VII,** Gotha, 1871.

Gauss, C. F. (1823), Theoria Combinationis Observationum Erroribus Minimis Obnoxiae Commentationes societatis regiae scientiarum Gottingensis recentiores, **V,** pp. 1–104, Gottingae; reprinted in Carl Friedrich Gauss Werke, Band **IV,** Göttingen, 1873.

Gauss, C. F. (1839), letter to F. W. Bessel dated February 28, 1839, reproduced in "Kritische bemerkungen zur methode der kleinsten quadrate," pp. 142–148 in Carl Friedrich Gauss Werke, Band **VIII,** (B. G. Teubner, Leipzig, 1900).

Gauss, C. F. (1857), Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections; English translation by Charles Henry Davis (Little, Brown and Co., Boston).

Gnedenko, B. V. (1962), The Theory of Probability (English translation by B. D. Sechler), (Chelsea Publishing Co., New York, N.Y.).

Hermach, F. L. (1961), An analysis of errors in the calibration of electric instruments, Communication and Electronics (AIEE) **54,** pp. 90–95.

Hillebrand, W. F., G. E. F. Lundell, H. A. Bright, J. I. Hoffman, Applied Inorganic Analysis, 2d ed. (1953), (John Wiley & Sons, Inc., New York, N.Y.).

Holman, Silas Whitcomb (1892), Discussion of the Precision of Measurements, (John Wiley and Sons, New York, N.Y.).

Keyser, Cassius J. (1922), Mathematical Philosophy, (E. P. Dutton and Co., New York, N.Y.).

Kline, S. J. and F. A. McClintock (1953), Describing uncertainties in single-sample experiments, Mech. Eng. **75,** pp. 3–8.

Laplace, Pierre Simon (1886), Theorie Analytique Des Probabilites; 3d edition, Vol. 7 of Oeuvres Completes de Laplace publiées sous les auspices de l'Académie des Sciences, (Gauthier-Villars, Imprimeur-Libraire de l'École Polytechnique, du Bureau des Longtudes, Successeur de Mallet-Bachelier, Quai des Grands-Augustins, 55, Paris).

McNish, A. G. and J. M. Cameron (1960), Propagation of error in a chain of standards, IRE Trans. Instr. **I-9,** pp. 101–104.

Millikan, R. A. (1903), Mechanics, Molecular Physics, and Heat, (Ginn and Co., New York, pp. 195–196).

Murphy, R. B. (1961), On the meaning of precision and accuracy, Materials Research and Standards **4,** pp. 264–267.

NPL (1957), Calibration of temperature measuring instruments, National Physical Laboratory Notes on Applied Science, No. 12, pp. 29–30, (Her Majesty's Stationery Office, London).

Ostle, Bernard (1954), Statistics in Research, (The Iowa State College Press, Ames, Iowa).

Parzen, Emanuel (1960), Modern Probability Theory and its Applications, (John Wiley & Sons, New York, N.Y.).

Proschan, Frank (1953), Confidence and tolerance intervals for the normal distribution, J. Am. Stat. Assoc. **48,** pp. 550–564.

Rossini, F. D. and W. Edwards Deming (1939), The assignment of uncertainties to the data of chemistry and physics, with specific recommendations for thermochemistry, J. Wash. Acad. Sci. **29,** pp. 416–441.

Schenck, Hilbert, Jr. (1961), Theories of Engineering Experimentation, (McGraw-Hill Book Co., Inc., New York, N.Y.).

Schrock, Edward M. (1950), Quality Control and Statistical Methods, (Reinhold Publishing Corp., New York 18, N.Y.).

Shewhart, W. A. (1931), Economic Control of Quality of Manufactured Product, (D. Van Nostrand Company, Inc., New York, N.Y.).

Shewhart, Walter A. (1939), Statistical Method from the Viewpoint of Quality Control, (The Graduate School, U.S. Department of Agriculture, Washington, D.C.).

Shewhart, Walter A. (1941), Contribution of statistics to the science of engineering, University of Pennsylvania Bicentennial Conference, Volume on Fluid Mechanics and Statistical Methods in Engineering, pp. 97–124, (University of Pennsylvania Press, Philadelphia).

Simon, Leslie E. (1941), An Engineer's Manual of Statistical Methods, (John Wiley & Sons, Inc., New York, N.Y.).

Simon, Leslie E. (1942), Application of statistical methods to ordnance engineering, J. Am. Stat. Assoc. **37,** pp. 313–324.

Simon, Leslie E. (1946), On the relation of instrumentation to quality control, Instruments **19,** pp. 654–656 (Nov. 1946); reprinted in Photogrammetric Engineering **XVIII,** pp. 566–573 (June 1952).

Spinks, A. W. and T. L. Zapf (1954), Precise comparison method of testing alternating-current watthour meters, J. Research NBS **53,** pp. 95–105.

Student (1908), The probable error of a mean, Biometrika **VI,** No. 1, pp. 1–25.

Student (1926), Mathematics and agronomy, Journal of the American Society of Agronomy **XVIII,** 703–719.

Student (1927), Errors of routine analysis, Biometrika, **XIX,** pp. 151–164.

Swindells, James F. (1959), Calibration of liquid-in-glass thermometers, NBS Circ. 600, pp. 11–12, (U.S. Government Printing Office, Washington 25, D.C.).

Tukey, J. W. (1960), Conclusions vs. decisions, Technometrics **2,** No. 4, pp. 423–433.

Waidner, C. W. and H. C. Dickinson (1907), On the standard scale of temperature in the interval 0° to 100° C, Bul. Bur. Stds. **3,** pp. 663–728.

Webster's Dictionary of Synonyms (1942, 1st ed.), (G. and C. Merriam Co., Springfield, Mass.).

Youden, W. J. (1950), Comparative tests in a single laboratory, ASTM Bulletin No. 166, pp. 48–51.

Youden, W. J. and J. M. Cameron (1950), Use of statistics to determine precision of test methods, Symposium on Application of Statistics, Special Technical Publication No. 103, pp. 27–34, (American Society for Testing Materials, Philadelphia).

Youden, W. J. (1951a), Statistical Methods for Chemists, (John Wiley & Sons, New York, N.Y.).
Youden, W. J. (1951b), Locating sources of variability in a process, Ind. Eng. Chem. **43,** pp. 2059–2062.
Youden, W. J. (1953), Sets of three measurements: The Scientific Monthly **LXXVII**, pp. 143–147.
Youden, W. J. (1954–1959), Statistical Design, Industrial and Engineering Chemistry. Feb. 1954 to Dec. 1959, Bimonthly articles collected in a single booklet available from Reprint Department, ACS Applied Publications, 1155 Sixteenth St., Washington 6, D.C.
Youden, W. J., W. S. Connor, and N. C. Severo (1959), Measurements made by matching with known standards, Technometrics **1,** pp. 101–109.
Youden, W. J. (1960), The sample, the procedure, and the laboratory, *Anal. Chem.* **32,** *pp. 23A–37A.*
Youden, W. J. (1961a) How to evaluate accuracy, Mat. Res. & Std. **1,** pp. 268–271.
Youden, W. J. (1961b), What is the best value? J. Wash. Acad. of Sci. **51,** pp. 95–97.
Youden, W. J. (1961c), Statistical problems arising in the establishment of physical standards, Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability **III,** pp. 321–335 (University of California Press, Berkeley and Los Angeles.).
Youden, W. J. (1961d), Systematic errors in physical constants, Phys. Today **14,** pp. 32–42 (1961); also in Technometrics **4,** pp. 111–123 (1962).
Youden, W. J. (1962a), Experimentation and Measurement, National Science Teachers Association Vistas of Science Series No. 2, (Scholastic Book Services, New York 36, N.Y.).
Youden, W. J. (1962b), Measurement Agreement Comparisons, presented at the Standards Laboratory Conference, National Bureau of Standards, Boulder, Colo., August 8–10, 1962.
Youden, W. J. (1962c), Uncertainties in calibration, IRE Trans. Instr. **I–11,** pp. 133–138 (1962).
Youden, W. J. (1962d), Realistic estimates of errors in measurements, ISA Journal **9,** No. 10, pp. 57–58.

(Paper 67C2-128)

On Absolute Measurement*

N. ERNEST DORSEY† AND CHURCHILL EISENHART‡

Dr. Dorsey did both his undergraduate and his graduate work at The Johns Hopkins University, receiving his A.B. in 1893 and his Ph.D. in Physics in 1897. After teaching at The Johns Hopkins unitl 1901, he joined the U S. Department of Agriculture. In 1903 he went to the Bureau of Standards as an assistant Physicist, later becoming Physicist until his retirement in 1943. He is a former associate editor of International Critical Tables, *and the author of numerous scientific papers and books. Dr. Eisenhart has served as Chief of the Statistical Engineering Laboratory of the National Bureau of Standards since 1946. He became a member of the faculty of the University of Wisconsin in 1937. During World War II he was engaged in mathematical research on aerial gunnery and rockets. Dr. Eisenhart is the author of more than 50 technical papers and is one of the editors and authors of the book* Selected Techniques of Statistical Analysis.

BY an absolute measurement of a physical quantity, such as the velocity of light, is meant the determination of the value of that quantity in terms of the significant fundamental units of length, mass, time, etc., and of those constant parameters that characterize the accepted system of theoretical equations that connect the several pertinent quantities. (p. 9.)

Theory of Errors

The mean of a family of measurements—of a number of measurements of a given quantity carried out by the same apparatus, procedure, and observer—approaches a definite value as the number of measurements is indefinitely increased. Otherwise, they could not properly be called measurements of a given quantity. In the theory of errors, this limiting mean is frequently called the "true" value, although it bears no necessary relation to the true quaesitum, to the actual value of the quantity that the observer desires to measure. This has often confused the unwary. Let us call it the limiting mean.

Let e denote the amount by which a given member of the family departs from the limiting mean, and let e_q denote that value which in the indefinitely extended family is surpassed by half of the e's; that is, it is an even chance that a given member of such an extended family departs from the limiting mean by as much as e_q.

The quantity e_q, the quartile error, commonly called the probable error of a single observation, will in this study be called the technical probable error of a single member of such a family. (p. 4.)

It should be noticed that the technical probable error either of a single measurement or of the mean of a group of n measurements indicates merely the closeness with which that measurement or mean probably approaches the limiting mean. It tells nothing whatever about the actual quaesitum, and so it is of very minor interest to the experimental physicist engaged in absolute measurements.

To him its main interest is threefold: (a) It tells him when it has become profitless to take additional routine observations; but in most cases other and more important considerations set another limit. (b) It may enable him to state positively that a systematic error affects one or both of two rival families of measurements. (c) It, as applied to a relatively small number of observations, enables him to state positively that systematic errors smaller than a certain amount cannot with certainty be detected experimentally with the apparatus and procedures employed in obtaining those measurements.

The last is, for him, by far the most valuable property of the technical probable error. But in practice he seldom thinks of it in that connection. By what seems to be a kind of intuition, he recognizes rough numerical relations between the mini-

* Excerpts from introductory "Remarks" of N. Ernest Dorsey's "The Velocity of Light" (*Transactions of the American Philosophical Society*, Vol. 34, pp. 1–110, 1944), selected and arranged by Churchill Eisenhart.

† Physicist (retired), National Bureau of Standards. Associate editor, *International Critical Tables*, 1922–1929.

‡ Chief, Statistical Engineering Laboratory, National Bureau of Standards.

On Absolute Measurement*

N. ERNEST DORSEY† AND CHURCHILL EISENHART‡

Dr. Dorsey did both his undergraduate and his graduate work at The Johns Hopkins University, receiving his A.B. in 1893 and his Ph.D. in Physics in 1897. After teaching at The Johns Hopkins unitl 1901, he joined the U S. Department of Agriculture. In 1903 he went to the Bureau of Standards as an assistant Physicist, later becoming Physicist until his retirement in 1943. He is a former associate editor of International Critical Tables, *and the author of numerous scientific papers and books. Dr. Eisenhart has served as Chief of the Statistical Engineering Laboratory of the National Bureau of Standards since 1946. He became a member of the faculty of the University of Wisconsin in 1937. During World War II he was engaged in mathematical research on aerial gunnery and rockets. Dr. Eisenhart is the author of more than 50 technical papers and is one of the editors and authors of the book* Selected Techniques of Statistical Analysis.

BY an absolute measurement of a physical quantity, such as the velocity of light, is meant the determination of the value of that quantity in terms of the significant fundamental units of length, mass, time, etc., and of those constant parameters that characterize the accepted system of theoretical equations that connect the several pertinent quantities. (p. 9.)

Theory of Errors

The mean of a family of measurements—of a number of measurements of a given quantity carried out by the same apparatus, procedure, and observer—approaches a definite value as the number of measurements is indefinitely increased. Otherwise, they could not properly be called measurements of a given quantity. In the theory of errors, this limiting mean is frequently called the "true" value, although it bears no necessary relation to the true quaesitum, to the actual value of the quantity that the observer desires to measure. This has often confused the unwary. Let us call it the limiting mean.

Let e denote the amount by which a given member of the family departs from the limiting mean, and let e_q denote that value which in the indefinitely extended family is surpassed by half of the e's; that is, it is an even chance that a given member of such an extended family departs from the limiting mean by as much as e_q.

The quantity e_q, the quartile error, commonly called the probable error of a single observation, will in this study be called the technical probable error of a single member of such a family. (p. 4.)

It should be noticed that the technical probable error either of a single measurement or of the mean of a group of n measurements indicates merely the closeness with which that measurement or mean probably approaches the limiting mean. It tells nothing whatever about the actual quaesitum, and so it is of very minor interest to the experimental physicist engaged in absolute measurements.

To him its main interest is threefold: (*a*) It tells him when it has become profitless to take additional routine observations; but in most cases other and more important considerations set another limit. (*b*) It may enable him to state positively that a systematic error affects one or both of two rival families of measurements. (*c*) It, as applied to a relatively small number of observations, enables him to state positively that systematic errors smaller than a certain amount cannot with certainty be detected experimentally with the apparatus and procedures employed in obtaining those measurements.

The last is, for him, by far the most valuable property of the technical probable error. But in practice he seldom thinks of it in that connection. By what seems to be a kind of intuition, he recognizes rough numerical relations between the mini-

* *Excerpts from introductory "Remarks" of N. Ernest Dorsey's "The Velocity of Light" (Transactions of the American Philosophical Society,* Vol. 34, pp. 1–110, 1944), selected and arranged by Churchill Eisenhart.

† Physicist (retired), National Bureau of Standards. Associate editor, *International Critical Tables,* 1922–1929.

‡ Chief, Statistical Engineering Laboratory, National Bureau of Standards.

mum detectable error and the mean deviation of the several determinations from their mean. And he studies those deviations without thinking about the technical probable error. Actually, the relations he uses are practically those that may be derived in the following manner from the technical probable error.

The argument runs as follows: If the means of two groups of measurements do not differ by at least the sum of their technical probable errors, then the existing difference is not sufficient to justify the assumption that they do not belong to the same statistical family. Consequently, if the only basic difference between the groups were the presence in one of a systematic error that was absent from the other, then the presence of that error could not be certainly established from the difference, unless it amounted to at least the sum of the two technical probable errors. Conversely, it cannot be proved that the measurements are not affected by such an error. (pp. 5 and 6.)

. . . the term "systematic error" is used to cover all those errors which cannot be regarded as fortuitous, as partaking of the nature of chance. They are characteristics of the system involved in the work; they may arise from errors in theory or in standards, from imperfections in the apparatus or in the observer, from false assumptions, etc. To them, the statistical theory of errors does not apply. They are frequently called "constant errors," and very often they are constant throughout a given set of determinations, but such constancy need not obtain. For example, if the value found by a certain measurement depends upon the humidity of the air, which the experimenter fails to record, thinking that it is of no consequence, then the measures will be affected by a systematic error which will, in general, vary throughout the day and especially from day to day. (p. 6.)

Averaging

Any set of numbers may be weighted as desired, and summed and averaged, and the result can be carried out to as many digits as one may wish. The procedures are simple, exact, and not open to any question or criticism. They are purely arithmetical.

But if the numbers represent physical quantities, then questions arise concerning both the validity of averaging and the number of digits that have a physical significance.

1) It is sometimes forgotten that the averaging of a set of values, even of the same kind, may be a physically invalid procedure. That is, that the average may not deserve greater confidence as an estimate of the quaesitum than do the individual values.

For example, consider a series of sets of determinations, each set being affected by a systematic error peculiar to it; that error being constant throughout any given set, but varying from set to set. Superposed on that error are fluctuating errors of various kinds. These last are minimized, set by set, by averaging the determinations composing a set. This averaging is entirely proper. But it leaves one with a series of values that differ, one from another, on account of the presence of systematic errors peculiar to each. In general, the averaging of such a series of values will be quite invalid; in general, the average will not deserve more confidence than do the individual values. The only cases in which it will be justifiable when the values differ by more than can be accounted for by the irregularities inherent in each of the several sets, are three: those in which it is definitely known—or perhaps is very highly probable—that the variation in the systematic error from one value to another either is (*a*) strictly fortuitous, in which case the fluctuating part of the error is minimized by the averaging, or (*b*) arises from the error fluctuating between equal and fixed positive and negative values, the number of positive values being essentially equal to the number of negative ones, or (*c*) arises from the error varying progressively from a positive value to a negative one as certain uncontrolled conditions change, and those conditions are known to vary in such a way that each negative error will in the long run be matched by an equal positive one.

Only when one knows a great deal about the systematic error can one be sure that any of these conditions are satisfied. And when he knows that much, he can often arrange to eliminate, or to evaluate, the error; and he should do so.

The cases that most frequently give trouble are those in which the data give evidence of the presence of a systematic error, but the experimenter does not know its source, and those in which another studying the data finds evidence of a systematic error that was overlooked by the experimenter. In such cases one may not know how the error varies with the conditions. If it makes all the values too great, then the smaller ones will be better than the average. Or the reverse may be true. Or the error may be present in some and absent from others; then averaging will not improve things.

Under such conditions it is quite improper to present the average as being superior to the individual values.

One is never justified in merely guessing that averaging will minimize or eliminate the effect of a systematic error. He must know it, must know that under the actually existing conditions the error is so minimized or eliminated.

In the absence of such knowledge, the proper brief summation of the work would seem to consist in a giving of the extreme values with a statement that at least some of the values seem to be affected by a systematic error of unknown origin. To this might well be added the experimenter's opinion, and if he wishes, the arithmetical average, with a clear statement of its questionable value. To give merely the average tends to mislead the reader, to blind him to the presence of systematic errors. The reader must always be on guard, as it is not very uncommon for a writer to average his results quite invalidly, either because he has not awaked to the fact that averaging may be invalid or because he has failed to recognize the evidence for the existence of systematic error.

2) The number of digits that are of physical significance in the sum and in the average must be carefully considered. (pp. 6 and 7.)

Quaesitum

The quaesitum of the investigation is the actual value of the quantity. The particular value yielded by a given apparatus, procedure, and observer is of no interest in itself, but only in connection with such a study as will enable one to say with some certainty that the value so found does not depart from the quaesitum by more than a certain stated amount. No investigation can establish a unique value for the quaesitum, but merely a range of values centered upon a unique value. The quaesitum may lie anywhere within that range, but the wiser and more careful the experimenter's search for systematic errors, and the more completely he has eliminated them, the less likely is it to lie near the limits of the range. The wider the range, the less becomes the physical significance of the particular value on which the range is centered. (p. 9.)

Definitive Value

The term "definitive value" is used in two distinct, though related, senses. (*a*) In a narrower, particular sense, it denotes the value that is believed to lie as near the quaesitum as any that can be legitimately derived from the observations taken in the course of the work being reported. It is the ultimate or definitive value to which that work itself leads. It is often called the "final" value of the work. (*b*) In a broader, general sense, it denotes the value that is believed to lie as near the quaesitum as any that can be derived from a consideration of all the determinations that have been made, and of all other available pertinent information. Whenever not otherwise indicated by the context or a modifier, it is in this broader sense that the term is to be understood.

Every report of measurements of a physical quantity should state clearly the particular definitive value to which those measurements lead. It may also give the broader definitive value based on everything that is known. But the two should not be confused, as unfortunately they often are. (p. 9.)

Dubiety

The determination of the range is of an importance that is secondary only to that of its center. No absolute measurement has been completed until values have been established for both of those quantities. The determination of the range necessarily involves an element of judgment, and the limits cannot be set with precision. Nevertheless, it is possible to assign a lower limit; and although no fixed upper limit can be assigned, it is possible to say that if suitable care and diligence had been employed, it is not likely that the range exceeds a certain specified value.

In order to distinguish this range from the numerous kinds of "errors" that abound, its half will in this study be called the "dubiety" of the value found. If that value be denoted by V, and the dubiety by D, then the quaesitum will likely lie within the range $(V - D)$ to $(V + D)$. By this, one means that nothing has come to light in the course of the work to indicate that the quaesitum lies outside that range.

The dubiety is made up of three distinct additive terms to which it is convenient to give descriptive names. They are as follows:

Mensural dubiety arises from the uncertainties in the several primary measurements and in the elimination of known systematic errors. It is common practice to take the arithmetical sum of the effects of these individual uncertainties as an upper limit for the mensural dubiety.

Discordance dubiety arises from the fact that the discordance in the individual determinations limits the smallness of a systematic error that can be experimentally detected. The result cannot be less dubious than the size of the largest systematic error that can escape detection. This term of the dubiety is generally the most important by far, and the least understood and least appreciated by those who are not experimentalists.

Deficiency dubiety arises from the determinations

being too few; in particular, finite in number. It is equal to the technical probable error of the result. This term, much honored by those not skilled in experimentation, is always smaller than the discordance dubiety and frequently is negligible in comparison therewith.

Of these three terms, the second alone needs to be especially considered here. In searching for systematic errors, the logical procedure is to make a series of measurements, then to change something and to make another series, and to compare the means of the two groups. This will be repeated as often as may seem necessary. None of the series can be long, for an extended delay offers opportunity for unanticipated changes to occur. If the two means being compared do not differ by more than the sum of their technical probable errors, their difference is of no physical significance—it proves nothing. Hence, the presence of a systematic error that does not exceed the sum of the technical probable errors of the two groups of observations used in the search cannot be established without great difficulty, if at all. That sets a minimum limit for the discordance dubiety. (pp. 9 and 10.)

Obviously, no one should claim a discordance dubiety that is smaller than the smallest systematic error that he might certainly have detected by the tests he made. But there may be reasons that seem to him sound for believing that the actual dubiety is smaller than that. In such case he may, and generally should, state his belief and the reasons therefor; but the statement should never be of such a kind as to lead the reader to confuse the writer's estimate with the minimum discordance dubiety as just defined. (p. 10.)

But on comparing a series of determinations made by different persons with significantly different apparatus and procedures, it may be found that the several members of the series agree more closely than their individual dubieties would lead one to expect. Then if the differences in apparatus and procedure are sufficiently fundamental, one might be justified in thinking it very improbable that the quaesitum lies far outside the range of the means of the several members of the series. And from the whole he might infer a smaller range of possible values than that demanded by the dubieties of the several determinations. (p. 10.)

No one is really interested in how near the quaesitum the definitive value may possibly lie, for he knows that by chance the two may coincide even though the work be very poorly done. But one does keenly desire to know how far the two are likely to differ—how dubious the definitive value may be. And it is the plain duty of the experimenter not merely to show that his definitive value may be that of the quaesitum, but to prove that it is unlikely to depart from the quaesitum by more than a certain stated amount. In order to obtain the information needed to meet that demand, the careful experienced investigator will proceed somewhat as follows. (p. 10.)

Procedure

Before one undertakes an absolute measurement in physics, he will make a careful theoretical study of the problem, including, among other things, methods of attack, sources of errors and how they can be avoided or eliminated, and types of computation. On the basis of that study, the apparatus will be constructed and set up. Only then does the investigation itself begin.

Working standards of the absolute units required must be carefully compared with primary standards. This will ordinarily be done at some standardizing laboratory, which will certify those working standards as being correct under certain specified conditions to within, say, a in 10^n. That value is accepted by the experimenter and sets the top limit to the known accuracy attainable in the work. If, for example, the absolute measurement attempted were simply a length, and the working standard were certified as correct to 3 in 10^5, then the absolute measurement (which determines merely the ratio of the measured length to that of the working standard) could under no condition give the value of the quaesitum to a known accuracy that exceeds 3 in 10^5. No matter how small the technical probable error of the measurements might be, the dubiety of the result cannot be less than 3 in 10^5. Indeed, the dubiety of the value found for the quaesitum will in general be distinctly greater than that, on account of errors inherent in the absolute measurement itself.

The experimenter will measure each of the involved quantities in terms of the appropriate working standard, taking pains to observe as well as may be the conditions laid down by the standardizing laboratory, and to determine carefully whatever is necessary to correct for the actual deviations from those conditions. He will do this repeatedly, and he will also measure them under deliberately different conditions, so as to obtain a check on the accuracy with which he can correct for departures from the specified conditions.

Having found that the apparatus seems to be

working properly, he will change, one by one, and by known amounts, each of the adjustments, and will note how each change affects his observations. If possible, he will carry each maladjustment to a point where it produces an easily measurable change in his observations; and if maladjustments in both directions (positive and negative) are possible, he will similarly study each. Thus he will find how important the several adjustments are, the accuracy with which they must be made, and perhaps how to detect each maladjustment experimentally and to correct for the error that it produces.

Readjusting the apparatus, he will proceed to change, one by one, every condition he can think of that seems by any chance likely to affect his result, and some that do not, in every case pushing the change well beyond any that seems at all likely to occur accidentally.

There still remains the possibility of systematic errors arising from unsuspected causes, from secular variation in laboratory conditions (temperature, humidity, light, vibration, etc.), possibly from solar, lunar, or atmospheric effects, etc. So the observer will take long series of observations, extending over weeks, months, or years, noting carefully everything that seems either pertinent in itself or of assistance in fixing the attendant conditions. These will be worked up, day by day, carefully compared with one another, and probably plotted in such a way as to show clearly any change that might appear. From time to time changes will appear, and will be studied.

Thus the experimenter presently will feel justified in saying that he feels, or believes, or is of the opinion, that his own work indicates that the quaesitum does not depart from his own definitive value by more than so-and-so, meaning thereby, since he makes no claim to omniscience, that he has found no reason for believing that the departure exceeds that amount.

That is exactly what he means. He does not mean as some have suggested, that he is of the opinion that the chances are only one in a hundred, or in a thousand, or in some other number n, that the quaesitum's departure from his definitive value exceeds that amount. He, differing from those others, feels that it would be foolish for him to make such a statement, that it could be nothing more than a gambler's guess. For how can one say, without stultifying himself, that the chance is one in n that the error produced in his result by an entirely unknown, and possibly non-existent, cause exceeds so-and-so, n being a definite specified number? And what can the word "chance" mean in that connection? Quantitative "chance" has significance only in relation to a family of events, and its value for a given event depends upon the characteristics of the family as well as upon that of the event itself. But as regards the uneliminated systematic errors, his observations define no family. He has nothing from which to compute a chance. All he can validly do is to express an opinion; and that opinion can validly relate only to certain theoretical considerations and to the magnitude of the errors that might have escaped his attention, not to any chance that his result might be in error by a given amount.

In every report, such an opinion of the limits within which the quaesitum is believed to lie, based solely on the work being reported, should be given. But in addition to that, previous measurements of the same quantity, when available, will usually be compared with those being reported, for one or more of the following purposes: supporting the author's value; setting other limits for the range within which the author thinks the quaesitum lies, deriving a general definitive value. But even in these cases only the same kind of opinion can be expressed, the number of absolute determinations that have been made of any given physical quantity being far too small to define a statistical family. (pp. 10 and 11.)

The experimenter's opinion must rest on evidence, if it is to have any weight. And the only evidence available comes from theory, the series of observations made in the course of the work, and the diligence with which errors were sought. These, and in particular the discordance of the observations and the diligence of the search, are what must be depended upon. Dependence on theory is weak, for the actual conditions never accord exactly with those assumed in the theoretical work.

He knows that it is impossible to avoid systematic errors, that even when he has done his best, his result is still haunted by the ghosts of such errors. His whole problem has been to seek such errors out, and to eliminate them when found; and he believes that in his long search any existing combination of them that would have produced an effect greater than the limit he sets would have been found. But he would be the first to admit that he may be wrong, that his result might be affected by a much larger error arising in such a way that, in spite of the many changes made in the course of the work, it remained essentially unchanged; but he thinks that contingency is highly unlikely. However, he is not entitled to that opinion unless he has carried out the indicated search, for in no other

way can a foundation be found on which to base an opinion.

In the absence of such a search, the worker can do no more than hope that all is going well. The fact that he sees no reason for suspecting the presence of an unknown systematic error is of no importance at all, no matter who the observer is. The really troublesome errors are exactly those that are not suspected. The suspected ones can usually be to some extent eliminated. (p. 12.)

Report

The work should be fully reported, so that the reader may know what was done, may have the means for forming an independent judgment of the work and for checking possible errors and omissions, and may have the worker's experience to build upon in case he himself should undertake a similar piece of work. The last is certainly a very important function of such a report, and should never be ignored.

The report should, of course, give a clear indication of the care with which search was made for sources of error, and of the thought that was given to it. Otherwise, one has no choice but to conclude either that no search was made, or that the author attached no special importance to it. In either case, the work is of little if any, objective value; its acceptance can rest only on authority, on subjective grounds.

Data should be reported as fully as may be. But in every series of observations some are erratic especially at the start. How should they be treated? Those that occur in the body of the work should certainly be reported as fully as if they were not erratic, and if the cause of the trouble is known, that should be explained.

Those that occur peculiarly at the beginning of the series, arising mainly from maladjustment and inexperience, furnish very valuable information regarding details of adjustment and manipulation that had escaped the foresight of the worker, and that might, therefore, readily escape the attention of the reader and of subsequent workers. In certain cases they give valuable information about unsuspected sources of error. For such reasons, they should never be completely omitted. They need not always be given in full, but they should be given to such an extent and in such detail as will show the reader what they were like and how they were related to the pertinent conditions, and should be accompanied by such explanatory text as will show him how they were regarded by the worker, and how he contrived to remove the disturbing conditions.

In brief, the report should give the reader a perfectly candid account of the work, with such descriptions and explanations as may be necessary to convey the worker's own understanding and interpretation of it. Anything short of that is unfair to the writer as well as to the reader. Every indication that significant information has been omitted reduces the reader's confidence in the work.

It is the unquestioned privilege of the worker to say where the boundary lies between preliminary or trial determinations, made primarily for studying and adjusting the apparatus and procedures, and those that were expected to be correct. But he should give good reasons for placing that boundary where he does; and those preliminary determinations should be reported to the extent already indicated.

Furthermore, it is scarcely fair, to anyone concerned, to describe a series of determinations as "preliminary," thus implying, in accordance with common usage, that they are open to question, that they are merely preparatory for something better, and then, later on, to include that same series in the list of good, acceptable determinations. To do so, both confuses the reader and suggests to him that the use of the adjective "preliminary" may have been merely a face-saving device intended to justify the ignoring of that series in case it should be found to disagree uncomfortably with later ones. (pp. 12 and 13.)

Miscellaneous

To say that an observer's results are influenced by his preconceived opinion does not in the least imply that those results were not obtained and published in entire good faith. It is merely a recognition of the fact that it seems more profitable to seek for error when a result seems to be erroneous, than when it seems to be approximately correct. Thus reasons are found for discarding or modifying results that do violence to the preconceived opinion, while those that accord with it go untested. An observer who thinks that he knows approximately what he should find labors under a severe handicap. His result is almost certain to err in such a direction as to approach the expected value.

The size of this unconsciously introduced error is, obviously, severely limited by the experimenter's data, by the spread of his values. The smaller the spread, the smaller, in general, will be this error. The size will be much affected also by the circumstances of the work, and by the strength of the bias. If the work is strictly exploratory, its primary

purpose being to find whether the procedure followed is at all workable, then only a low accuracy will be expected, and there will be no serious attempt to explain departures from the expected, even though the departures be great. Consequently, this error of bias may be entirely absent from such results. But if the worker is striving for accuracy, then departures from the expected will appear to him serious; and the stronger the bias, the more serious will they seem. He will seek to explain them; and that seeking will tend, in the manner already stated, to introduce an error. An error arising in this way will seldom be negligible, but in no case should one expect it to be great, the work being done in good faith. (p. 2.)

. . . published definitive values, with their accompanying limits of uncertainty, are not experimental data, but merely the authors' inferences from such data. Inferences are always subject to question; they may be criticized, reexamined, and revised at any time. (p. 3.)

. . . it is every author's duty to publish amply sufficient primary data and information to enable a reader to form a just and independent estimate of the confidence that may be placed in the inferences that the author has drawn therefrom. If he does not, he is false to both his reader and himself, and his inferences should carry little weight, no matter how great his reputation may be; . . . .

Indeed, values reported without such satisfactorily supporting evidence have no objective value whatever, no matter how accurate they may happen to be. They rest solely on the authority of the reporter, who is never infallible. (p. 3.)

Made in the United States of America
Reprinted from THE SCIENTIFIC MONTHLY
Vol. LXXVII, No. 2, August, 1953

SYSTEMATIC ERRORS in PHYSICAL

The author is a consultant to the National Bureau of Standards on the statistical and mathematical design of experiments in physics, chemistry, and engineering. Dr. Youden joined NBS in 1948.

By W. J. Youden

PHYSICISTS today make very little use of statistical techniques. There was good reason for the minor role so long accorded the statistical evaluation of the errors in physical constants. When two laboratories make independent determinations, each may attach to its "best" value a ± sign followed by an estimate s of the error. This estimate of the error is often based upon a series of observations made under carefully controlled conditions. Experimenters soon discovered that if laboratories A and B reported values C_A and C_B for the same constant, the difference Δ between C_A and C_B was almost always a large multiple of the estimated error s_a (or s_b). Obviously these calculated errors had no more to do with the real errors than the neatness of the laboratory or the promptness with which the investigator answered his mail.

Statisticians in turn sensed that all the observations made in one laboratory, with one piece of equipment, were afflicted with some fairly constant and unknown increment that was a resultant of biases associated with the method of measurement, with the particular assembly of apparatus, and perhaps with some more or less persistent characteristics of the environment. The statistician saw no way either to detect or to assess these "constant" errors. Consequently, statisticians concentrated on other activities where random errors were all that really mattered. The comparison of the yields obtained from two or more varieties of wheat involves only comparisons. Similarly the chemist, seeking to find for an industrial process a set of operating conditions that will give maximum yield, or maximum profit, can compare runs and not worry much that all the results may be half a percent high. That may be discovered later, when the annual inventory is taken.

Both physicists and statisticians apparently agreed to part company. There remained the custom of calculating and reporting the precision of the measurements, partly to establish that very precise habits of work were maintained, and partly in the hope that more weight would be given to a determination if a very small precision error was attached to the result. All recognized that a small precision error was necessary but gave no guarantee that the reported average was close to the truth.

For decades there has been but little contact between experimental physics and statistics, and I think that both parties have been the losers for giving up so easily. Statisticians were not aware that many of the physical measurements either approximate, almost exactly, certain ideal statistical models or else suggest the invention of statistical models that would extend statistical theory. The physicist, in turn, relying on his experimental skill, continued to track down the sources of his errors by traditional methods and overlooked certain advantageous ways of combining his observations.

This paper discusses three main topics. First some remarks will be made regarding the statistical confidence limits that apply to two or three independent determinations of a constant. The major section deals with what appears to be a plausible explanation for the unexpectedly large differences between the values obtained in different laboratories. The last portion presents some statistical aids for tracking down the causes for disagreement among laboratories.

Independent Determinations of a Constant

SUPPOSE laboratories A and B report the values C_a and C_b for the same constant. Precision estimates s_a and s_b may or may not be given. Perhaps the investigators have searched their souls and ventured to indicate the likely maximum errors in the values reported. These estimated errors generally do not determine the opinions of other laboratories regarding these two results. Depending on the laboratory visited, you may encounter one of four possible opinions:

1. The laboratory favors C_a and discounts C_b
2. The laboratory favors C_b and discounts C_a
3. The laboratory believes C_a and C_b are of about equal merit
4. The laboratory is a sceptic and believes both C_a and C_b unreliable.

If the laboratories are approximately split between the first two opinions and one of the determinations is close to correct, then the obvious statistical conclusion can be drawn that about half of the laboratories will eventually be disappointed. Perhaps all will be disappointed if neither determination is near the correct value.

If most of the laboratories are of the fourth opinion, clearly there is no statistical problem. But if a majority

CONSTANTS

of the laboratories feel that both results are worthy and that there is little to choose between the two determinations, then some statistical remarks may be made. We are going to suppose that the method of measurement is a new one and consequently there is no other information available than that contained in the two results already in hand. That there will be some difference between the two results is to be expected. Examination of the difference between the two results tells us little because we have no way of knowing whether this difference is smaller or larger than usual. Statistical tables show that if the average difference between duplicates is ten units, then individual differences of from one to thirty units are not uncommon. So a single difference may be very misleading.

Suppose a third laboratory is about to make a report. If we assume that the three results are independent of the order in which they were obtained, some simple logic suggests that there is a one-third chance that the last result reported will be intermediate in magnitude between the first two results reported. Denote the smallest, middle, and largest results by s, m, and l. These three letters can be arranged in six orders: *sml*, *slm*, *msl*, *mls*, *lsm*, and *lms*. For two of these six sequences the middle result m is the last in the sequence. Consequently, without ever knowing the first two results, it is a fair gamble to bet one to two that the third result will lie between the first two results.

Notice, too, that this logic holds quite apart from any knowledge as to how closely the first two agree. Of course, if by chance the first two values are identical or nearly so, one might argue that it would be less likely to get a third result between them than if the first two did not agree closely. But just what other standards can one produce to say, in any particular case, what would be close agreement, or what would constitute poor agreement, if these two results constitute all the information available?

A closely similar question, given two equally esteemed results, is: What is the chance that the two values C_a and C_b bracket the correct value? The answer is one half. After all, the correct result does not go gallivanting around the way a third independent result might and contributes no error. It is quite remarkable that this conclusion rests on a very modest assumption about the underlying distribution, of which these two constitute our sole information. We have only to concede that if a goodly number of qualified laboratories undertook to make determinations, that about half the determinations would be smaller and the remainder larger than the correct value. Symmetry is all that is required. So it is a coin-tossing problem with two coins where heads refer to plus deviations and tails to minus deviations. A quarter of the time we get two heads (both results high), a quarter of the time two tails (both results low), and half the time a head and a tail, or deviations of unlike sign which means that the results bracket the correct value.

I hasten to admit it is conceivable, through some defect in theory, that all the results are afflicted with a component error of the same sign and this will spoil our coin-tossing game. But this is speculation and tantamount to saying that it is useless *ever* to venture an opinion about the confidence to be placed in the determinations. It does seem appropriate to be aware of the probabilities that I have given even if one cautiously states the assumptions upon which the probabilities are calculated.

Now a probability of one half is not a very comforting figure and it is a natural thing to wonder how we might extend our thinking to limits outside the two reported values in order to attain a greater confidence that the correct value lies within these limits. Let $C_a - C_b = \Delta$, where $C_a > C_b$, and suppose we consider limits of the following kind:

$$\text{Upper limit} = C_a + k\Delta, \quad \text{Lower limit} = C_b - k\Delta.$$

It now becomes necessary to examine how sensitive our confidence is to the kind of distribution that would fit a collection of such determinations. Suppose we assume first the traditional normal distribution. Then for k equal to one, the probability is about 0.8; that is, adding the difference between two results to the larger one, and subtracting it from the smaller, gives limits that four times out of five should bracket the correct value. If instead of the normal distribution, we imagine that a determination is equally likely to fall anywhere within some finite, but unknown, interval centered on the unknown correct value, the probability drops from 0.80 to 0.75. And there is a vast difference between the bell-shaped normal distribution and the "rectangular" distribution of equal probability for all values over a finite range.

Table 1 shows, for the normal distribution, how the probability of bracketing the correct value between $C_a + k\Delta$ and $C_b - k\Delta$ increases with k. Remember that Δ is the difference between two determinations that are accorded equal weight.

Table 1. Probability, P, that $C_a+k\Delta$ and $C_b-k\Delta$ bracket the correct value. Normal distribution assumed. Equal weight accorded C_a and C_b; $\Delta=C_a-C_b$.

k	0	1	2	3	4	5	6	7
P	0.5000	0.795	0.874	0.910	0.930	0.942	0.951	0.958

Sad to say, it takes an over-all spread between the upper and lower limits of 13 times the difference between the two determinations to attain the traditional 95-percent confidence limits. You may reply: "Nonsense. Things are not that bad." But you should be prepared to justify your comment. After all, in the light of the peripatetic wanderings of the "accepted" value of some of our constants, how can you, from just two determinations, form a better judgment about the correct value?

The real explanation of the wide limits required in Table 1 is the small amount of information we have on Δ. One pair may give a Δ considerably smaller or considerably larger than the average Δ if many such pairs were available. The way to improve matters is to get additional, truly independent determinations.

The gain in assurance that comes from a third independent determination at first seems disproportionately large. The narrowing of the confidence limits comes not so much from being able to average three rather than two, but from having a firmer grip on the extent of agreement that may be expected among independent determinations. The chance that the three results bracket the correct value rises from one half to three quarters. That is, the chance that both a tail and a head will be obtained when three coins are tossed is six out of eight. If the difference between the largest and smallest of the three values is added to the largest value and subtracted from the smallest value, we obtain confidence limits that have slightly better than a twenty to one chance of bracketing the correct value. Of course the same assumptions discussed above for two determinations are made here too. Even if there are grave reservations about these assumptions, one can say that the chances are no better than those indicated. A bound has been set to our optimism.

In spite of the difficulties that arise in estimating the error in a constant most scientists agree that the effort should be made. Professor Bridgeman in his talk at the 1960 Gordon Conference on "Information Processing for Critical Tables of Scientific Data",[1] emphasized that critical tables should endeavor to present the "best" value and to make some estimate of the "probable" error of the value selected.

Composite Character of Systematic Errors

SOMETIMES successive measurements may be made in a time interval so short that it is reasonable to regard the measurements as being made with no changes in environment, apparatus, or any other condition that might affect the measurement. Given adequate precision a reasonable number of measurements serve to establish an average that very closely characterizes the measuring system during this interval. This average will differ more or less from the correct value. This departure from the correct value is plainly the algebraic sum of several small effects. For example, the diameter of a diaphram, the resistance of a coil, the temperature and volume of a chamber, and similar quantities will all be assigned values that depart in some degree from the actual values that existed while the measurements were made. These deviations from the actual values influence the outcome—and each may either add or subtract some small increment to the measurements. The experimenter has surely tried to keep these various increments somewhat the same in size and usually he would say it was a toss up as to the sign of each increment.

As an example, a recent determination of g presents two sets of 32 measurements—one set made with one rule, the other set with a second rule.[2] Fig. 1, taken from this paper, shows the distribution of the measurements for each set. No elaborate statistical test is required to make convincing the reality of the difference between the means of the two sets. Doubtless there were other components or conditions that had similar increments.

Imagine ten such increments of about equal magnitude but unpredictable in sign. Now the experimenter is surely at the mercy of the laws of chance. There are six different algebraic sums (each either plus or minus) depending on how fate has grouped the signs of these increments.

Division of the signs	*Algebraic sum*	*Frequency*	
5 and 5	0	252	
4 and 6	± 2	420	912
3 and 7	± 4	240	
2 and 8	± 6	90	
1 and 9	± 8	20	112
0 and 10	± 10	2	
		1024	

The foregoing tabulation shows that about once out of nine times the increments gang up on the helpless experimenter and introduce a composite systematic error at least six times as large as the small "uncertainty" he has achieved in his values for the components in his apparatus. There is a chance in three of a net sum of four or more increments. If the experiment is repeated in another laboratory, the same situation holds and half the time the two composite net sums will be of opposite sign. We now see how the difference between the results from the two laboratories may be an order of magnitude greater than the standard of accuracy set for the individual components.

The individual increments are taken as equal in size to simplify the presentation. If the increments vary from small to large, the effect is very nearly the same if their average magnitude equals the "standard" increment used above. While there is a certain amount of cancellation because there may be both plus and minus increments, it is the net *sum* that matters. There is no averaging out here. So the distribution of these "sums" depends on the average size and the number of contributing increments. Experimenters properly enough

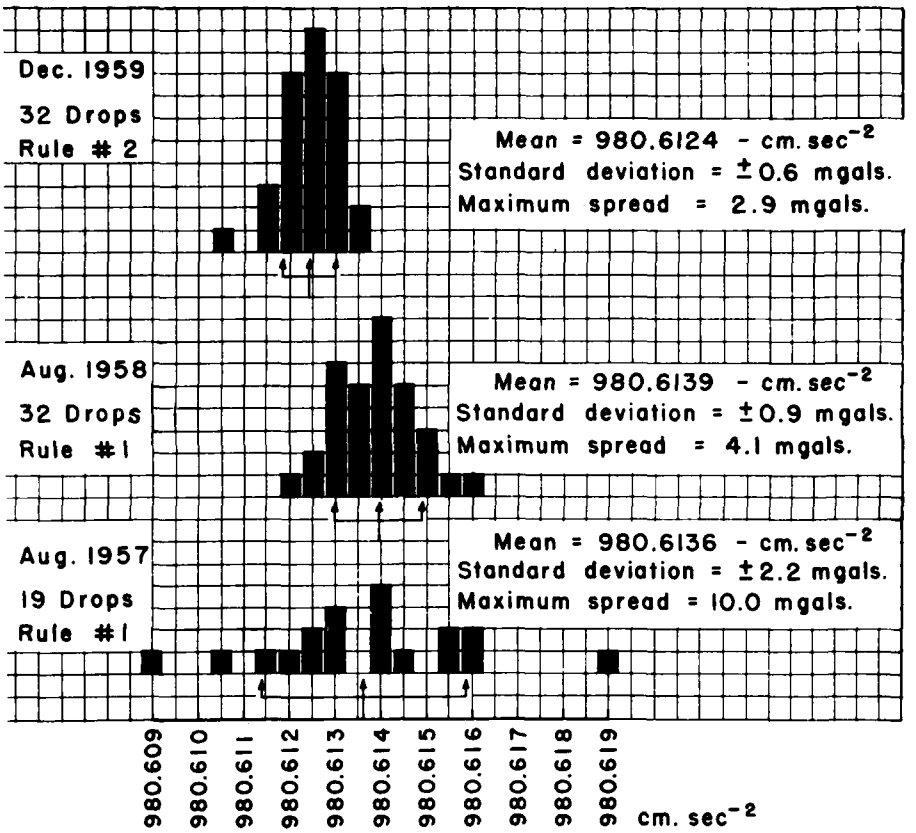


Fig. 1. Measurement of gravity constant

direct their best efforts to the detection and reduction of the larger increments because this is the most effective way to reduce the average size of the increments.

Detection of Increments of Systematic Errors

WE have seen that an aggregate of systematic errors, all of them individually relatively small, can nevertheless sum up in such a fashion as to produce a substantial net displacement from the correct result. The detection of small systematic errors, and by that I mean errors comparable to the precision error, requires a considerable number of repeat measurements. Fig. 1 shows *32* repeat measurements of the gravitation constant g with each of two different rules. The repeat measurements with a rule cluster around a central value for that rule and offer convincing evidence that there is a real difference between the averages for these two rules. The shape of the scatter of the measurements around their average is what would be expected on the basis of the normal distribution of errors. Suppose the difference Δ between the averages for the two rules is equal to s, the standard deviation of the repeat measurements. Then, reference to tables for the normal distribution shows that it is necessary to make at least eight repeat measurements on each rule before we can conclude, with 95% confidence limits, that the rules differ at all in their mean values.

The important thing here is, that within one laboratory, the *precision* measure of error is the proper measure to use in evaluating differential effects of such substitutions of components of the apparatus, or in evaluating effects of changing environmental conditions. Dorsey, in a lengthy paper published in 1944,[3] gives on pages 10 and 11 some pointed remarks on the necessity of examining the effect of changing the adjustments of the apparatus. I quote one sentence.

> Readjusting the apparatus, he (the experimenter) will proceed to change, one by one, every condition he can think of that seems by any chance likely to affect his result, and some that do not, in every case pushing the change well beyond any that seems at all likely to occur accidentally.

Excerpts from Dorsey's 110-page article are given in a paper by Dorsey and Eisenhart.[4]

The single sentence quoted above is particularly interesting because Dorsey saw the direction in which progress was to be made. In the nearly twenty years since Dorsey prepared his remarks we have made considerable progress in the direction he indicated. We see that not only should the adjustments be changed, but whenever possible there should be at least duplicate components for certain vital parts of the apparatus. The use of two rules, as exhibited in Fig. 1, shows how much the results are at the mercy of a single rule. Clearly the only thing to do is to take the average for

the *two* rules, and there are *only* two rules. The prediction as to what might happen with more rules throws us right back to the discussion in the first part of this paper. Incidentally, Dorsey's recommendation that substantial changes be introduced in the conditions indicates that he found it difficult to detect the effects of small changes.

The previously-quoted sentence contains the phrase, "one by one". Change the adjustments "one by one" is the way we all learned to experiment. The interpretation is easy then because, for example, if we merely substitute one rule for another, any effect is obviously to be credited to the substitution of one rule for the other. In the intervening years since Dorsey wrote there has been a good deal of activity in the devising of more efficient programs for evaluating the effect of just such changes in adjustments or substitutions of components in the apparatus.

If there are a number of possible adjustments and components to investigate, the total number of measurements may become very large because a considerable number of repeat measurements must be made for each assembly and each adjustment. There are really two parts to this problem. If, for example, the experimenter winnows his choices down to seven alternatives (including both adjustments and substitutions for components) does that mean that he need try all 2^7, or 128, possible combinations? Experimenters have already answered this question. They designate some standard initial assembly and set of adjustments and then proceed to change, one by one, the seven items under consideration. Some measurements are made under the initial state; an item is changed, and another set of measurements made. Whatever was changed is put back to the initial state and a second item changed. There will be eight such sets and a goodly number of measurements are required in each set.

Today, as a result of some purely theoretical inquiries into what statisticians term *weighing designs*, we know that seven variables could have been equally well evaluated with one fourth the usual number of measurements. Or, and the prospect is enticing, we could have detected, and perhaps corrected, systematic effects only half as large as those just detectable under the "one by one" approach. I say that these were theoretical statistical inquiries because statisticians were mainly concerned with biological and chemical problems that involved *major* changes in the variables. In such investigations there are mutual interactions of the variables that pose quite different problems. Here the changes in the variables are minute. The differential effect of substituting one rule for another almost identical rule (as in Fig. 1) would be virtually unaltered even if some other set of initial conditions had been chosen.

Statisticians were unaware of the extremely important problems posed in the evaluation of physical constants. Yates was the first statistician to suggest and name "weighing designs" in an incidental paragraph in a paper [5] in 1935. In fact, Yates belittled the designs because he deemed it most unlikely that any problems appropriate for such designs really existed. It is interesting that other statisticians,[6-10] in a purely theoretical way, embellished the idea advanced by Yates. None saw the possibilities that exist for application in very precise physical measurements. And we lack, even now, an adequate exploration of the programs that might serve the needs of those who determine physical constants.

Once it is recognized that the effect of a very small change in a variable does not depend on the other variables, provided that these other variables also are held within very close limits, the way is open to change more than one variable at a time. I illustrate this principle first for the case of three variables x, y, and z, which may be assigned other nearby values x', y', and z'. Let us designate the standard initial condition by x, y, and z and let these serve as the coordinates of the origin in the three-dimensional graph shown in Fig. 2.

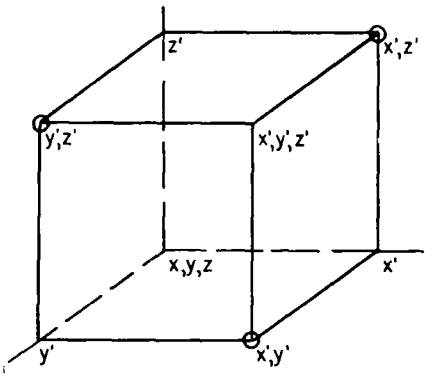


Fig. 2. Diagram for three variables

The customary way to explore this situation is to change one variable at a time. The three choices are to move to x' on the x axis, to y' on the y axis, and to z' on the z axis. These are poor choices by comparison with the choices $x'y'$, $x'z'$, and $y'z'$—marked with circles in the diagram.

The usual procedure for detecting the effect of changing x to x' makes use of the data obtained at the two points x,y,z and $x'yz$. The more efficient method for detecting the effect of changing x to x' makes use of the data obtained at all four points, x,y,z; $x,y'z'$; $x'y'z$ and $x'yz'$. Two of these sets involve x and two involve x' so the data are grouped accordingly.

xyz	$x'y'z$
$xy'z'$	$x'yz'$

The two sets with x include y and z and y' and z'. So the average value for x incorporates the effects associated with y, y', z and z'. This is also visibly true for the two sets with x'. Therefore, the effect of changing x to x' will be given by comparing the *average* of all

the data taken at x with the *average* of all the data taken at x'. Inspection discloses that the four sets may be partitioned into appropriate pairs to detect the effect of changing y to y', or z to z'.

xyz	$xy'z'$	xyz	$xy'z'$
$x'yz'$	$x'y'z$	$x'y'z$	$x'yz'$

The basic idea here is so important that I illustrate it again for the case of just two variables. Recall the 32 measurements made with rule 1 (r) and the additional 32 measurements made with rule 2 (R). Group them in opposing groups as next shown.

$rrrr$	$RRRR$
$rrrr$	$RRRR$
$rrrr$	$RRRR$
$rrrr$	$RRRR$
$rrrr$	$RRRR$
$rrrr$	$RRRR$
$rrrr$	$RRRR$
$rrrr$	$RRRR$

If half of the measurements in each group were made with another variable at s and the remainder at S, the measurements may be segregated into four sets.

Set 1		*Set 2*		*Set 3*		*Set 4*	
rs	rs	rS	rS	RS	RS	Rs	Rs
rs	rs	rS	rS	RS	RS	Rs	Rs
rs	rs	rS	rS	RS	RS	Rs	Rs
rs	rs	rS	rS	RS	RS	Rs	Rs
rs	rs	rS	rS	RS	RS	Rs	Rs
rs	rs	rS	rS	RS	RS	Rs	Rs
rs	rs	rS	rS	RS	RS	Rs	Rs
rs	rs	rS	rS	RS	RS	Rs	Rs

Rule r is present in sets 1 and 2 and rule R in sets 3 and 4. The other variable is put at s in sets 1 and 4 and at S in sets 2 and 3. We may now play both ends against the middle pairs of sets and evaluate the effects of s and S. The data are used twice over. If the set size is reduced from 16 to 8, 7 variables may be studied with these same 64 measurements.

That is, *all* the data taken are used to evaluate the effect of changing each variable. Either fewer repeat measurements are required at each combination, or more variables may be investigated with the same number of measurements. Indeed, the more variables that are investigated in this manner, the more efficient this method becomes. Seven variables lend themselves to an especially elegant sequence of seven partitions of eight sets into contrasting sets of four sets against four sets. This example, shown in Table 2, I am glad to report, is the one first mentioned by Yates twenty-five years ago. You may note that four of the initial conditions are changed each time.

It would be a pleasure indeed if I could include here a small catalog of programs extensive enough to meet the situations likely to occur in practice. I can point out that the program shown in Table 2 can be used for fewer than seven variables by ignoring one or more of the changes. It will still be necessary to work with eight different combinations. Similarly, in a paper by Plackett and Burman [10] schemes for 12, 16, and more variables are given. Variables may be ignored here, too, but the number of combinations is not reduced.

The minimum number of combinations required is one more than the number of variables, if just two alternatives are used for each variable. The substitution of a component is sometimes a tedious affair so there is sure to be interest in programs involving a minimum number of combinations. I have tried my hand at this game and offer the program shown in Table 3 for studying five variables with six combinations. Each effect is measured using the results of four of the six combinations, divided two against two.

Table 2. Program for seven variables with eight sets.

1	2	3	4	5	6	7	8
t	t	t	t	T	T	T	T
u	u	U	U	u	u	U	U
v	V	v	V	v	V	v	V
w	w	W	W	W	W	w	w
x	X	x	X	X	x	X	x
y	Y	Y	y	y	Y	Y	y
z	Z	Z	z	Z	z	z	Z

Table 3. Five variables in six sets.

1	2	3	4	5	6	
v	V	v	V	v	V	$v-V=(1+5)/2-(2+6)/2$
w	w	W	W	W	W	$w-W=(1+2)/2-(5+6)/2$
x	X	X	x	X	x	$x-X=(1+4)/2-(2+3)/2$
y	Y	y	Y	Y	y	$y-Y=(3+6)/2-(4+5)/2$
z	z	Z	Z	z	z	$z-Z=(5+6)/2-(3+4)/2$
						av$=(1+2+3+4)/4$

Table 4. Program for three variables, two with three choices, one with two choices.

Variables

x	$\mathbf{x}$	X
y	$\mathbf{y}$	Y
$\mathbf{z}$		Z

Six sets

1	2	3	4	5	6
x	x	$\mathbf{x}$	$\mathbf{x}$	X	X
y	$\mathbf{y}$	y	Y	$\mathbf{y}$	Y
$\mathbf{z}$	Z	Z	z	z	Z

	x	$\mathbf{x}$	X
y	$\mathbf{z}$ 2	Z -2	
$\mathbf{y}$	Z 1		$\mathbf{z}$ -1
Y		$\mathbf{z}$ -1	Z 1

Above coefficients are weighing factors to estimate x-$\mathbf{x}$

There will be times when more than two choices are possible and of interest for some of the variables. I regret to say that the enumeration of efficient designs for such mixtures of two and three choices has hardly begun. Let me illustrate with a simple case of three choices for each of two variables and two choices for a third variable. There are $2 \times 3 \times 3$ possible combinations, and a minimum of six sets are necessary to sepa-

rate the individual effects of these variables. The problem is to pick that subset of six from the eighteen available sets that will lead to the most efficient evaluation of the effects of the variables. I suspect that the program shown in Table 4 is as good as any, just on the basis of the appealing symmetry.

In the squares are indicated certain factors and these are the factors to be used in evaluating the effect of changing x to $\mathbf{x}$. Notice that the estimation of the effect of a change in the variable involves a weighted average of the six set results. The best average for the constant gives equal weight to all six sets. Similar sets of constants apply for evaluating $\mathbf{x} - X$, $y - Y$, etc.

The essential point regarding these illustrative programs is that certain combinations lend themselves to an efficient use of the data, that is, to a more sensitive scrutiny of the possible sources of error. The one-at-a-time technique is one of the least efficient programs. The small individual contributions to error that are associated with uncertainties in values assigned to component quantities are not easy to detect. A planned set of combinations will rank the various sources of error in order of magnitude and reveal where the program is weakest. Statistical techniques will not remove errors but they can help in isolating the important sources of error.

Enduring Values

THERE is more in this discussion than the matter of efficiency. The several variables, chosen by the experimenter because they may influence the result, are actually put to the test. At present the investigator has two ways to arrive at an opinion or guess as to the error introduced by any one of the quantities which he would like to know exactly when he introduces it into his computations. He may, on his judgment, hazard a guess as to the maximum uncertainty in each of the relevant quantities. Alternatively, he may accept the estimates of others—e.g., the estimate of the man who measured the length of the rules used in the determination of g. Thermometers, weights, resistances, purities, standard cells—the list is endless—they may all be obtained with some sort of statement from the calibrating source. It is easy to push responsibility off this way. And we go on getting determinations from different laboratories that disagree much more than anticipated even when the claimed uncertainties in the components are included. Maybe it is time to check these indispensable bits of information. If the Coast and Geodetic Survey measures the distance employed in a determination of the velocity of light—ask them to measure two or three distances. The above schemes will soon put these measurements to the test. A choice of resistances, diaphragm diameters, thermometers—all should be made to run the gauntlet.

Yes, I know, all the resistances *may* be subject to the same bias. The two rules used in the determination of g *may* share a common increment that will not be revealed by the data. But there *is* a difference between the two rules and now our estimate of the limits of error can allow for this difference. We must use more than one rule, or we will not have the data to estimate this source of error.

I return to the summation of the systematic errors associated with the individual components. The use of two or more choices creates the possibility that the choices differ in the signs of their systematic errors. The final value reported will be an *average* of the results obtained with the several sets—each set a unique combination of components and conditions. The individual summations of the separate sets now enter into an average with all the advantages that come from taking an average. Furthermore, the spread of the results for the several sets will surely give a more realistic idea of the uncertainty in the final result than that obtained from hopeful guesses.

There is another matter that cannot be glossed over. Suppose the measurements are made according to some carefully thought out program similar to the suggested weighing designs. Admittedly this limits the freedom of the investigator. The experimenter likes to be free to follow some inspired hunch. He often wants to try some alteration in the apparatus, or in the conditions, on the chance that his spontaneous idea has merit. This might be regarded as the art rather than the science of experimentation. The investigator should consider how often such ideas pay off and also the large number of measurements required to detect small effects, when tempted by such ideas.

I personally hold that allowance should be made for "shots in the dark". If the planned program is allotted, say, around three quarters of the measurement time, there would still be opportunity for imaginative excursions. Even if these isolated shots lack the power that they would have if incorporated in the planned program, they add a lot of zest to experimentation.

We all know that a serious effort to determine a physical constant is not undertaken lightly. The dominating thought in the mind of the investigator is to arrive at an enduring value. What is an enduring value? I suggest that it is a result coupled with a stated zone of uncertainty that includes the value that future workers will converge upon.

References

1. P. W. Bridgman, "Critique of critical tables", Proc. Nat. Acad. Sci. **46**, 1394 (1960).
2. H. Preston-Thomas, L. G. Turnbull, E. Green, T. M. Dauphiness, and S. N. Kalra, "An absolute measurement of the acceleration due to gravity", Ottawa Division of Applied Physics, National Research Council, Ottawa, Canada.
3. N. E. Dorsey, "The velocity of light", Trans. Am. Phil. Soc. **34**, 1 (1944).
4. N. E. Dorsey and C. Eisenhart, "On absolute measurements", Sci. Monthly **77**, 103 (1953).
5. F. Yates, "Complex experiments", Suppl. J. Roy. Stat. Soc. **2**, 181 (1935).
6. H. Hotelling, "Some problems in weighing and other experimental techniques", Ann. Math. Stat. **15**, 297 (1944).
7. O. Kempthorne, "The factorial approach to the weighing problem", Ann. Math. Stat. **19**, 238 (1948).
8. K. Kishen, "On the design of experiments for weighing and making other types of measurements", Ann. Math. Stat. **16**, 294 (1945).
9. A. M. Mood, "On Hotelling's weighing problem", Ann. Math. Stat. **17**, 432 (1946).
10. R. I. Plackett and J. P. Burman, "The design of optimum multifactorial experiments", Biometrika **33**, 305 (1946).

Uncertainties in Calibration*

W. J. YOUDEN†

Summary—**This paper presents some methods for making comparisons between standards and items undergoing calibration. These methods may be used in a variety of measurements. The purpose is to accumulate data that provide objective estimates of the precision and that are also useful in detecting sources of systematic errors. This purpose is achieved in using some standard statistical designs in the scheduling of the work program.**

The problems of stating the uncertainty and of combining the uncertainties in a chain of calibrations are discussed.

Introduction

A LABORATORY that provides a calibration service soon seeks the answers to three questions. These questions all concern the uncertainty in the value assigned to an item that has been calibrated. The first question is usually directed to the magnitude of the uncertainty. Another question deals with the kind of data needed to support any claim made regarding the uncertainty. Finally the laboratory seeks some way of stating the uncertainty which will convey useful information to those who will make use of the calibration service.

One of the important points to be clear about is that any statement of the uncertainty applies to a class of closely similar items that are calibrated by a specified procedure with a particular assembly of apparatus. No one can say, for any individual item, whether the random error in the calibration is smaller or larger than the average error for the class of items. It is possible to evaluate the average error (or some other measure) for the calibration procedure as used on a group of similar items. If a systematic error has to be provided for, this error will also be carried over to all the items in the class.

This paper is primarily directed to those laboratories that undertake calibrations that involve comparisons of the test items with a "standard." The standard is an item that has a certificate from a national or other recognized laboratory. The certificate states the correction to be applied to the nominal value of the item. The standard item should be as nearly as possible the same magnitude as the test items. The game is simply one of comparing each test item in turn with the standard item using a suitable assembly of apparatus. This shows very clearly that the calibration is a comparison procedure. Very often, by one or another ingenious technique, comparisons can be made virtually bias-free. For example, if the standard weight is balanced against a dummy weight on the other pan of the balance, and then the test item substituted for the standard weight, the effect of inequality of balance arms is automatically eliminated. Also an appropriate alternation of repeated weights will cancel out drifts that may arise from environmental shifts. Success in devising a bias-free comparison makes the observed difference between the standard and test item subject to random errors only. The task then is to determine the error in the comparison procedure as used over a sequence of test items from the same class.

Although the discussion is intended for laboratories providing calibration services, some of it is directly applicable to the work of national laboratories that are the source of the standards used by calibrating laboratories. In some instances, as for example the kilogram, where an object is arbitrarily assigned a nominal value, the standards laboratory need consider only comparison errors. Many standards require "absolute" determinations—problems that challenge every resource of the experimental scientist.[1,2] Even here, once a value has been established, and an uncertainty assigned, the standard laboratory accepts this value for comparison activities. At this stage again, the comparison error becomes a matter of interest.

Determination of the Uncertainty Associated with a Specified Comparison Procedure

There are many ways of obtaining an estimate of the error in the difference, or in the ratio, of the two magnitudes associated with the comparison of two items. The direct repetition of measurements is the simple approach that involves the least computation. Simple repetition is vulnerable to "memory" on the part of the operator. Also there is often a failure to provide the opportunity for errors to manifest themselves. For example, the differences in EMF between two Weston cells may be determined with a potentiometer. Even if the potentiometer is considerably offset from the null point and a new null point found, this does not constitute a real repetition. Surely the cells should be disconnected and the connections remade, and the adjustment for the standard cell offset and reset. In fact all the operations should be repeated anew. Generally there is a temptation to slur over such tedious and time-consuming operations, perhaps to omit them entirely. Quite plainly if some of these operations do contribute to the error of a comparison, no mere repeating of the null point can possibly disclose the presence of such contributions to the error of a comparison. Yet such contributions do matter because they are present in the steps involved in using the apparatus.

* Received August 15, 1962. Presented at the 1962 International Conference on Precision Electromagnetic Measurements as Paper No. 3.1.

† National Bureau of Standards, Washington, D. C.

[1] A. G. McNish and J. M. Cameron, "Propagation of error in a chain of standards," IRE Trans. on Instrumentation, vol. I-9, pp. 101–104; September, 1960.

[2] W. J. Youden, "Comparative tests in a single laboratory," *ASTM Bull.*, pp. 48–52; May, 1950.

Generally it is much better to devise some indirect way of measuring the error of a comparison. Preferably the indirect way should make it impossible or quite difficult for the operator to have any idea what his error is as he makes his reading. Again it is emphatically better to base the estimate of the error on data accumulated in small sets over a number of test items and over a considerable period of time—possibly months. An estimate of error based on many readings on one test item and over a short time interval may fail to provide a representative sample of the conditions that do prevail over long periods and with many items. Some gage blocks may be more perfectly faced than others. The bore of a capillary will vary from item to item and various other characteristics may make some items better performers than others. Ascertaining the errors individually for each test item is quite impossible because of the vast amount of work this would require.

The endless variety of combinations of apparatus used for making comparisons between items means that the situation largely determines the type of indirect approach to the measurement of the error appropriate for the comparison. For this reason only some techniques of fairly general application will be discussed. The examples will serve to illustrate the indirect approach. Many comparisons may be best studied using an arrangement of the measurements specially devised for each individual problem.

One recurrent type of situation is characterized by the direct pairings of the items. The difference, or ratio, of the compared items is the quantity actually measured. Typical of this situation is the connecting of two Weston cells in opposition and the measurement of the net voltage provided by the two cells when so connected. The "pairing" of the cells is a physical, not a paper, transaction. Similarly the difference between two gage blocks may be measured when the blocks are adjacent to each other.

The basis for an indirect estimate is, in this case, the three pairings that can be made among three items. The usual trio would be made up of a standard and two test items. Even more desirable is a trio made up of two standards and one test item. In the latter case, two important advantages accrue. First, the value for the test item is tied to the average of two standards with the reduction of the error ascribable to uncertainty in the value for the standard. Second, evidence gradually accumulates on the experimental difference between the two standards. This quantity, when compared with the difference expected on the basis of the entries entered on the certificates for the standards provides a useful check on the national laboratory that issued the certificates. This information would be valuable to the national laboratory as a measure of its own performance to be set against whatever claims were made on their certificates.

Suppose then that the three items are S_1, T_1 and T_2. No change in the argument is needed if the items are S_1, S_2 and T_1. The three available comparisons are

$$S_1 - T_1 = d_1$$
$$T_1 - T_2 = d_2$$
$$T_2 - S_1 = d_3$$
$$(S_1 + T_1 + T_2) - (S_1 + T_1 + T_2) = d_1 + d_2 + d_3 = \Delta \to 0.$$

The three observed quantities, d_1, d_2 and d_3, will all be different because no two items are alike. The operator, therefore, is under no compulsion to get "checks" on repeat readings. Once the three differences are in hand, the tabulation just listed shows that the three differences, with properly chosen signs, should sum to zero if the measurements were without error. The sum will differ from zero because there are errors in the three measurements. The conclusion is, in consequence, that Δ, the amount by which the sum of the d's differs from zero, is a measure of the error of the comparisons.

Given a collection of Δ's from a sequence of trios, it is easy to calculate the experimental error of these comparisons. Let k trios be measured and let the resulting Δ's be $\Delta_1, \Delta_2, \cdots, \Delta_k$. The estimate of the standard deviation for an observation d_i is given by $s_d = \sqrt{(\Sigma\Delta^2/3k)}$. The number of degrees of freedom for this estimate, s_d, is k.

Consider the first choice—one standard and two test items. Let us see how to calculate an estimate of the difference between T_1 and S_1. We have the direct comparison which gives $T_1 - S_1 = -d_1$. There is also the indirect, and *independent* estimate obtained by adding d_2 and d_3. When the last two comparisons are added, T_2 drops out of the sum. So another estimate for $T_1 - S_1$ is $d_2 + d_3$. This additional information should not be ignored. What is the proper weight to give to each estimate? Double weight is given to the direct estimate, $-d_1$, and single weight to $(d_2 + d_3)$ because it is the *sum* of two measurements.

$$\text{The weighted estimate for } \overline{T_1 - S_1} = \frac{-2d_1 + d_2 + d_3}{3}.$$

$$\text{The weighted estimate for } \overline{T_2 - S_1} = \frac{2d_3 - d_1 - d_2}{3}.$$

$$\text{The weighted estimate for } \overline{T_1 - T_2} = \frac{2d_2 - d_1 - d_3}{3}.$$

The standard deviation for the above estimates is given by $\sqrt{2/3}\, s_d$.

One of the happy consequences of these improved estimates is that the new estimates are consistent. Notice that the sum $\overline{(T_1 - S_1)} + \overline{(S_1 - T_2)} + \overline{(T_2 - T_1)}$ equals zero as it should. The other important source of information is Δ. Once a sequence of Δ's is available, a very important channel of information has been opened. In theory each Δ should be equally likely to be plus or minus. The opportunity exists to incorporate in these differences some interesting feature of the apparatus

used for the comparison. Thus, if the symbols represent standard cells, then each cell may have been directly connected once to the left terminal of the potentiometer and once to the right terminal. If the symbols represent meter bars, the arrangement would be made as follows:

Position in the comparison chamber

Right end		Left end		
S_1	−	T_1	=	d_1
T_1	−	T_2	=	d_2
T_2	−	S_1	=	d_3

$$(S_1+T_1+T_2)-(S_1+T_1+T_2)=d_1+d_2+d_3=\Delta$$

Now Δ includes the difference between the right and left ends of the comparison chamber. Should there be an undetected or uncorrected persistent temperature difference between the two ends of the chamber, the bars would be longer when placed in the warmer end. This would tend to make the Δ's for a succession of triad comparisons, predominantly of one sign. Thus the Δ's may be utilized to detect the presence of a systematic error which ought not to be present in comparison procedures.

Even if there is an undue predominance of Δ's with the same sign, the data are still informative about the error of the comparison procedure. The average Δ gives an estimate of three times the effect associated with position in the chamber. Fortunately the arrangement of the items has cancelled out this effect on the comparison because each test bar was once in each end of the chamber. That is to say, the estimates of the *differences* between the bars have not been biased. The Δ's that would have been observed if there were no bias may be obtained by subtracting from each Δ the average of the Δ's. The corrected Δ's may be designated by Δ''s. These Δ's should now split about evenly between plus and minus. The standard deviation for the comparison process, *i.e.*, for any single measured difference d_i between two bars, is

$$s_d = \sqrt{\frac{\Sigma\Delta'^2}{3(k-1)}}$$

with $(k-1)$ degrees of freedom. The number of triads available is given the symbol k. The question of whether the average Δ differs enough from zero to constitute important evidence for a bias may be judged by the quantity

$$\sqrt{\frac{\Sigma\Delta'^2}{k(k-1)}}.$$

If the average Δ exceeds a stated multiple (two or somewhat more depending on the number of Δ's available) of this last expression, the evidence suggests a bias. If the evidence for a bias is not substantial, the standard deviation for a single measurement is calculated from the Δ's by the previously given formula $s_d=\sqrt{(\Sigma\Delta^2/3k)}$. The number of degrees of freedom should be twenty or more but preliminary estimates of the standard deviation may be made with a few degrees of freedom.

The device just illustrated may be generalized to cope with comparisons involving quintets or more. Five items make available ten pairings as follow:

Left	Right		Left	Right
A	B		A	C
B	C		C	E
C	D	and	E	B
D	E		B	D
E	A		D	A

The pairs may be grouped in two sets each of which provides one estimate of a left vs right effect. Five degrees of freedom are available per ten pairings for the estimation of the standard deviation.

All of the above discussion may be recast if the comparison measurement gives the ratio of the values for the two items. If $A/B=x$, $B/C=y$ and $C/A=z$, the product xyz of the three ratios would be exactly unity if the measurements were made without error. The amount by which the product xyz differs from unity is a measure of the errors in the measurements. The argument follows the same line as before. The weighted estimates are

$$A/B=\sqrt[3]{x^2/yz},\quad B/C=\sqrt[3]{y^2/xz}\quad\text{and}\quad C/A=\sqrt[3]{z^2/xy}.$$

Another way to accumulate evidence of the errors in calibrations may be illustrated by the calibration of platinum resistance thermometers. If the calibrating laboratory possess the facilities to set up the silver, gold or other calibrating mediums, the appropriate equation may be fitted to the data. An arbitrary bath temperature may then be set up and its temperature measured with a thermometer calibrated by a national laboratory and also measured by the test thermometer just calibrated. The difference, Δ, between these two readings is a measure of the performance of the calibrating laboratory. Again a predominance of one sign among the Δ's indicates a bias relative to the national laboratory. The Δ's should be accumulated over a number of test thermometers.

If the standard thermometer itself is used to establish the temperatures used to calibrate the test thermometer the above procedure is useless for the detection of bias but still useful as a measure of the random errors. Every care must be taken, for example, that the resistances are accurately measured by checking the equipment against known resistances of similar magnitude.

Another avenue of approach to the error of comparisons is illustrated by angle blocks. Here matters can usually be arranged so that closure should result, *i.e.*, the angles measured should sum up to 360°. Enough has

been said to suggest that the general principle of indirect estimates of the errors of comparison may take on many forms in actual practice. The important point to note is that the error estimated in this manner is more likely to approach the real error than an estimate based on many repetitions of a simple comparison.

How Should the Uncertainty be Stated?

The discussion thus far has dealt with two questions simultaneously: the estimate of the error and the evidence to support the estimate. Oddly enough it seems an even more difficult task to come to widespread agreement on how to record the errors.

Return to the first problem considered. A calibrating laboratory possesses a standard unit with a certificate issued by a national laboratory. The laboratory also possesses the necessary equipment for making comparisons. Furthermore a sufficient number of test items have undergone comparison with the standard so that a good estimate is available of the error of the comparison procedure. We assume a negligible systematic error. Let s be the estimate of the standard deviation for the comparison. What should the calibrating laboratory put on any certificate it issues to its customers? Is it enough to put some generally accepted multiple of s as an indication of the maximum error likely to have been made in the comparison? Not quite it seems, because the certificate issued by the national laboratory also has on it a statement about the uncertainty in the value entered on its certificate.

Weight and length standards offer as simple a case as possible. The uncertainty entered on the certificate issued by the standard laboratory involves a sequence of comparisons. The first comparison is that of its own reference standard A with an object I that, by definition, is one kilogram or one meter. Succeeding steps relate the reference standard A to subordinate or working standards.

Let us suppose that a national prototype meter bar A is compared with the international standard, I. (Actually the standard for length is now based on the wavelength of light, but for some purposes comparisons between bars are still made.) The object is to determine a correction C to be applied to the nominal value of the national standard. The work of comparison will also provide an estimate s for the standard appropriate for the correction based on the repeated readings. Let C_t be the true value for the correction and C_0 be the observed correction. If the comparison has been achieved without introducing a bias, the expectation is that C_0 does not differ from C_t by more than a small multiple of s. The multiple for s is a personal choice and is usually in the range from two to three.

At this point it is well to pause and consider meter bar A. The observed correction C_0 is almost certainly not equal to the true correction C_t. The difference (C_0-C_t) is not known but is a physical, unchanging magnitude.[1] No matter how many other bars are compared with standard A, this unchanging error is carried over into every calibration. This error (C_0-C_t) is not now a random quantity. It is fixed by two quantities. The true difference between A and the international standard I is C_t and we hope that this difference is constant. The other quantity C_0 is the experimentally determined estimate of C_t and this will not be changed until the national standard A is again taken to France. Both C_0 and C_t are constants and the difference between them must therefore be a constant. To pretend that this difference varies in successive uses of the standard for calibration is sheer nonsense. True, out of two score national standards each with a similar error (C_0-C_t), there will be some where (C_0-C_t) is nearly zero, and perhaps two countries whose (C_0-C_t) exceeds $2s$. Any one country takes a substantial risk if it assumes that it possesses a correction that is very close to the true correction. In the absence of any guide, the safe thing is to guard against the worst that may reasonably happen to the estimate of the correction for meter bar A.

There is a redeeming feature about this situation. In practice, a series of intermediate standards such as I, A, B, C, is interposed between I and some item D sent in for calibration. We may argue that, for each step, we can envisage an unknown error (C_0-C_t).

Thus,

comparison $A-I$ is in error by $(C_0-C_t)_A$,
comparison $B-A$ is in error by $(C_0-C_t)_B$,
comparison $C-B$ is in error by $(C_0-C_t)_C$,
comparison $D-C$ is in error by $(C_0-C_t)_D$.

Now each of these errors is a constant but the signs of the errors are as likely to be plus as minus. To the extent that the signs are not all the same, there will be compensation and reduction in the over-all sum of these errors. With four errors, there is a one-in-eight chance that all the errors have the same sign. Even if these errors all have the same sign, there is only a very remote chance that every one of them is large.[1]

We wish to establish for item D a maximum error that has some specified small chance of being exceeded. The four errors were obtained by drawing one random error from each of four populations with standard deviations s_I, s_A, s_B and s_C. The error in D is therefore found by the usual quadrature formula

$$s_D = \sqrt{s_I^2 + s_A^2 + s_B^2 + s_C^2}.$$

The uncertainty which the national laboratory records on the certificate for D will be some multiple of s_D.[1]

The calibrating laboratory that gets item D and its certificate has a different problem. Suppose this laboratory could make its comparisons without any error. The

[1] C_0-C_t is practically a constant for the best meter bars. Electrical meters, bridges, resistances, etc., may undergo some slow drift with time or continued use.

laboratory would then copy onto the certificates that it issues the uncertainty given on the certificate for D. In fact, this laboratory, when it calibrates items E_1, E_2, E_3, $\cdots$, does have a comparison error s_E. One possibility is to combine s_D and s_E by quadrature obtaining $\sqrt{s_D^2+s_E^2}$ as the combined standard deviation. This distribution will be centered on C_0 and is shown as the dotted curve in Fig. 1. Unfortunately the actual situation is different. The random errors are in fact distributed with standard deviations s_E around the unknown true value C_t for the standard D. The solid curve in Fig. 1 shows this situation where C_0-C_t is, in the illustration, 1.6 s_D. The vertical dashed line has been drawn at $2\sqrt{s_D^2+s_E^2}$ to the right of C_0. This leads to the expectation that 2.5 per cent of the errors will be in the area to the right of the vertical line. The actual distribution for the chosen value of C_0-C_t puts 15 per cent of the area to the right of the dashed line. Of course if the calibrating laboratory has a standard deviation, three or four times as large as (C_0-C_t) is ever likely to be, then the effect of the off-center positioning becomes negligible.

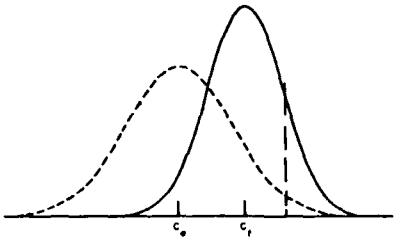


Fig. 1—Comparison of distribution centered on true value (solid line) with distribution centered on experimental value (dotted line).

The important conclusion from the above remarks is that the actual error in the value on a certificate may be a *random* error from the viewpoint of the national laboratory but it is a *systematic* error for the calibrating laboratory. The calibrating laboratory can use only the value ascribed to its standard by the national laboratory. The value is used over and over. A lucky calibrating laboratory will get a standard with a small random error and, in consequence, a small systematic error is introduced in its comparisons. Other laboratories not so fortunate will have a much larger error. But who is lucky and who is unlucky is not known so the safe thing to do is for all calibrating laboratories to act as though they had the maximum possible *systematic* error that might arise from the uncertainty in the value entered on the certificate that applies to their standard.

How shall such a systematic error be merged with the random error associated with the calibrations it undertakes? The answer depends on the use that will be made of these calibrated items. If they go to lower echelons to serve as "standards" the argument used for the sequence of meter bars appears appropriate and the two errors are merged by quadrature. If the calibrated item is to be put to work making measurements routinely, the quadrature combination seems inappropriate. Here the systematic error may be added to the maximum likely random error in order to give an upper bound to the possible error in the measurements.

Most of the standards possessed by a national laboratory are known to have a systematic as well as a random component of error. It is a fair hazard that in most cases the systematic component is the dominant component. In that event, the uncertainty entered on the certificates issued by the standards laboratory may be only slightly larger than the systematic error. Regardless of the relative contribution of the two components, the calibrating laboratory must view the stated uncertainty as a systematic error. To gamble and attempt to view even a part of the uncertainty as a random component may be quite misleading. If the standard in the possession of a laboratory happened to have one of the larger random errors, then it may result that practically *all* of the certificates it issues will have an error in excess of the claimed error. Of course some other calibrating laboratory that happened to have a standard that in fact had a very small error will issue certificates that are all well within the claimed error. Over the whole country things will average out. But the customers of the unlucky calibrating laboratory will all be unlucky *as a group*.

The foregoing shows that if any one calibrating laboratory combines errors on the presumption of a random component in the uncertainty on its certificate, it forces upon *all* of its customers the effect of *one* random error. The laboratory has probably no intention of doing this. The laboratory does intend to make a statement that actually protects nearly all of its customers, and even the exceptions should be borderline.

Discussion

Consider the chain of comparisons starting with a national standard and continuing on down to some piece of equipment that will be used routinely. It appears usual to form an estimate of the error by studying each link in the chain. It does seem likely that in the later links of the chain, the equipment and the environment used for the comparisons will be definitely inferior to the facilities available for the initial comparisons. Thus, at some stage, the comparison uncertainty may be large in comparison with preceding uncertainties. It is easy to demand incredibly small uncertainties but evidence should be presented that these small uncertainties are really needed.

The author is of the opinion that the ultimate users of calibrated items often have an optimistic notion of the quality of the measurements they make. It is most unlikely that adequate studies have been made of the errors in the measurements made by ultimate users. It

was pointed out earlier that readings repeated in rapid sequence with little or no disturbance of the equipment cannot be expected to reveal the real errors. For example, a gage block may be used to "set" a very sensitive gaging device. The user compares two items and is impressed by the sensitivity of the device. A demand is then made for better standards.

What is called for is a little ingenuity to devise means that will disclose the real magnitude of the errors in the measurements. Closely similar test blocks should be periodically resubmitted for measurement. Near equality of the test blocks is needed to make identification difficult. The schedule of tests should be prepared in some artful or random sequence.

The following example[2] illustrates the point that operators calibrating clinical thermometers could "repeat" their readings extremely well even when the repeat reading followed the reading of 23 other thermometers. The *average* difference between repeat readings was 0.0102°—one twentieth of a scale division. The average of 24 readings should have an error of about $0.01/\sqrt{24}$ or 0.002. The holder and the 24 thermometers were set aside for a few days and then reread. Again the superb agreement between closely repeated readings was observed. Unfortunately the *average* of the readings shifted by more than 0.02°, *i.e.*, an order of magnitude greater than the expected error. This phenomenon was demonstrated over and over again with different operators and different sets of thermometers. Operators shifted relative to one another as much as 0.04° when their averages on the same thermometers were compared. There is little doubt that even more startling contrasts between real and apparent errors would appear when the environment itself has to be carefully controlled to minimize errors. Some investigations should be made in order to ascertain whether some of the demands made for better standards are really justified.

Conclusion

At every stage in the hierarchy of calibrating laboratories, there is a laboratory that has on a bench three objects of interest:

1) A standard item from the echelon above with a certificate,
2) An assembly of equipment appropriate for comparing items, and
3) A collection of items awaiting calibration for the echelon below.

One safe procedure for all calibrating laboratories would be to quote the uncertainty in its standard, to state the uncertainty in its comparison process, and to tell its customers that the simple sum of these two components is the only safe measure of the possible error in the value assigned to the item just calibrated.

It should not be overlooked that the uncertainty stated on the certificate accompaning the standard sometimes includes a "stability" allowance. On the basis of broad experience a reasonable estimate of the drift effects can be made. When the uncertainty assigned to the standard includes such an allowance, this information should also be given.

Reprinted from IRE TRANSACTIONS
ON *INSTRUMENTATION*
Volume I-11, Numbers 3 and 4, December, 1962
PRINTED IN THE U.S.A.

Expression of the Uncertainties of Final Results

Clear statements of the uncertainties of reported values are needed for their critical evaluation.

Churchill Eisenhart

Measurement of some property of a thing in practice always takes the form of a sequence of steps or operations that yield as an end result a number that serves to represent the amount or quantity of some particular property of a thing—a number that indicates how much of this property the thing has, for someone to use for a specific purpose. The end result may be the outcome of a single reading of an instrument, with or without corrections for departures from prescribed conditions. More often it is some kind of average, for example, the arithmetic mean of a number of independent determinations of the same magnitude, or the final result of a least squares "reduction" of measurements of a number of different magnitudes that bear known relations with one another in accordance with a definite experimental plan. In general, the purpose for which the answer is needed determines the precision or accuracy required and ordinarily also the method of measurement employed.

Although the accuracy required of a reported value depends primarily on the *intended* use, or uses, of the value, one should not ignore the requirements of other uses to which it is likely to be put. A reported value whose accuracy is entirely unknown is worthless.

Strictly speaking, the actual *error* of a reported value, that is the magnitude and sign of its deviation from the truth (*1*), is usually unknowable. Limits to this error, however, can usually be inferred—with some risk of being incorrect—from the precision of the measurement process by which the reported value was obtained, and from reasonable limits to the possible bias of the measurement process. The *bias*, or *systematic error*, of a measurement process is the magnitude and direction of its tendency to measure something other than what was intended; its *precision* refers to the typical closeness together of successive independent measurements of a single magnitude generated by repeated applications of the process under specified conditions; and its *accuracy* is determined by the closeness to the true value characteristic of such measurements.

Precision and accuracy are inherent characteristics of the measurement process employed and not of the particular end result obtained. From experience with a particular measurement process and knowledge of its sensitivity to uncontrolled factors, one can often place reasonable bounds on its likely systematic error (bias). It is also necessary to know how well the particular value in hand is likely to agree with other values that the same measurement process might have provided in this instance, or might yield on remeasurement of the same magnitude on another occasion. Such information is provided by the estimated *standard error* (*2*) of the reported value, which measures (or is an index of) the characteristic disagreement of repeated determinations of the same quantity by the same method, and thus serves to indicate the precision (strictly, the imprecision) of the reported value (*3*).

Four Distinct Forms of Expression Needed

The uncertainty of a reported value is indicated by stating credible limits to its likely inaccuracy. No single form of expression for these limits is universally satisfactory. In fact, different forms of expression are recommended, which will depend on the relative magnitudes of the imprecision and likely bias, and their relative importance in relation to the intended use of the reported value, as well as to other possible uses to which it may be put (*4*).

Four distinct cases need to be recognized: (i) both systematic error and imprecision negligible, in relation to the requirements of the intended and likely uses of the result; (ii) systematic error not negligible, imprecision negligible; (iii) neither systematic error nor imprecision negligible; and (iv) systematic error negligible, imprecision not negligible.

Specific recommendations with respect to each of these cases are made below. General guidelines upon which these specific recommendations are based are discussed in the following paragraphs.

Perils of Shorthand Expressions

Final results and their respective uncertainties should be reported in sentence form whenever possible. The shorthand form "$a \pm b$" should be avoided in abstracts and summaries; and never used without explicit explanation of its connotation. If no explanation is given, many persons will take $\pm b$ to signify bounds to the inaccuracy of a. Others may assume that b is the "standard error," or the "probable error," of a, and hence the uncertainty of a is at least $\pm 3b$, or $\pm 4b$, respectively. Still others may take b to be an indication merely of the imprecision of the individual measurements, that is, to be the "standard deviation," or the "average deviation," or the "probable error" of a single observation. Each of these interpretations reflects a practice of which instances can be found in current scientific literature. As a step in the direction of reducing this current confusion, it is recommended that the use of "$a \pm b$" in presenting results be limited to that sanctioned for the case of tabular results in the fourth recommendation of the section below headed "Systematic error not negligible, imprecision negligible."

The author is a senior research fellow and former chief of the Statistical Engineering Laboratory at the National Bureau of Standards, Washington, D.C. 20234. The recommendations presented in this paper have evolved at the Bureau over a period of many years and are made public here for general information, and to educe comments and suggestions.

Imprecision and Systematic Error Require Separate Treatment

Since imprecision and systematic error are distinctly different components of inaccuracy, and are subject to different treatments and interpretations in usage, two numerics respectively expressing the imprecision and bounds to the systematic error of the reported result should be used whenever both of these errors are factors requiring consideration. Such instances are discussed in the section below for the case of "Neither systematic error nor imprecision negligible."

In quoting a reported value and its associated uncertainty from the literature, the interpretation of the uncertainty quoted should be stated if given by the author. If the interpretation is not known, a remark to this effect is in order. This practice may induce authors to use more explicit formulations of their statements of uncertainty.

Standard Deviation and Standard Error

The terms *standard deviation* and *standard error* should be reserved to denote the canonical values for the measurement process, based on considerable recent experience with the measurement process or processes involved. When there is insufficient recent experience, an estimate of the standard error (standard deviation) must of necessity be computed by recognized statistical procedures from the same measurements as the reported value itself. To avoid possible misunderstanding, in such cases, the term "computed (or estimated) standard error" ("computed standard deviation") should be used. A formula for calculating this computed standard error is given in the section below for the case of "Neither systematic error nor imprecision negligible."

Uncertainties of Accepted Values of Fundamental Constants or Primary Standards

If the uncertainty in the accepted value of a national primary standard or of some fundamental constant of nature (for example, in the volt as maintained at the National Bureau of Standards, or in the acceleration of gravity g on the Potsdam basis) is an important source of systematic error affecting the measurement process, no allowance for possible systematic error from this source should be included ordinarily in evaluating overall bounds to the systematic error of the measurement process. Since the error concerned, whatever it is, affects all results obtained by the method of measurement involved, to include an allowance for this error would be to make everybody's results appear unduly inaccurate relative to each other. In such instances one should state: (i) that measurements obtained by the process concerned are expressed in terms of the volt (or the kilogram, or other unit) "as maintained at the National Bureau of Standards," or (ii) that the indicated bounds to the systematic error of the process are exclusive of the uncertainty of the stated value adopted for some particular constant or quantity. An example of the latter form of statement is:

> . . . neglecting the uncertainty of the value 6.6256×10^{-34} joule seconds adopted for Planck's constant.

Systematic Error and Imprecision Both Negligible

In this case the reported result should be given, after rounding, to the number of significant figures consistent with the accuracy requirements of the situation, together with an explicit statement of its accuracy. An example is:

> . . . the wavelengths of the principal visible lines of mercury-198 have been measured relative to the 6057.802106 Å (angstrom units) line of krypton-98, and their values in vacuum are
>
> 5792.2685 Å
> 5771.1984 Å
> 5462.2706 Å
> 4359.5625 Å
> 4047.7146 Å
>
> correct to eight significant figures.

It needs to be emphasized that if no statement of accuracy or precision accompanies a reported number, then, in accordance with the usual conventions governing rounding, this number will ordinarily be interpreted as being accurate within $\pm\frac{1}{2}$ unit in the last significant figure given; that is, it will be understood that its inaccuracy before rounding was less than $\pm$ 5 units in the next place. The statement "correct to eight significant figures" is included explicitly in the foregoing example, rather than left to be understood in order to forestall any concern that an explicit statement of lesser accuracy was inadvertently omitted.

Systematic Error Not Negligible, Imprecision Negligible

When the imprecision of a result is negligible, but the inherent systematic error of the measurement process concerned is not negligible, then the following rules are recommended:

1) Qualification of a reported result should be limited to a single quasi-absolute type of statement that places bounds on its inaccuracy.

2) These bounds should be stated to no more than two significant figures.

3) The reported result itself should be given (that is, rounded) to the last place affected by the stated bounds (unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that it may possess for certain particular uses).

4) Accuracy statements should be given in sentence form in all cases, except when a number of results of different accuracies are presented, for example, in tabular arrangement. If it is necessary or desirable to indicate the respective accuracies of a number of results, the results should be given in the form $a \pm b$ (or $a \pm^{b}_{c}$, if necessary) with an appropriate explanatory remark (as a footnote to the table, or incorporated in the accompanying text) to the effect that the $\pm b$, or $\pm^{b}_{c}$, signify bounds to the systematic errors to which the a's may be subject.

5) The fact that the imprecision is negligible should be stated explicitly.

The particular form of the quasi-absolute type of statement employed in a given instance will depend ordinarily on personal taste, experience, current and past practice in the field of activity concerned, and so forth. Some examples of good practice are:

> . . . is (are) not in error by more than 1 part in (x).
> . . . is (are) accurate within $\pm$ (x units) [or $\pm$ (x) percent].
> . . . is (are) believed accurate within (. . . . .).

Positive wording, as in the first two of these quasi-absolute statements, is appropriate only when the stated bounds to the possible inaccuracy of the reported value are themselves reliably established. However, when the indicated bounds are somewhat conjectural, it is desirable to signify this fact (and put the reader on guard) by inclusion of some modifying expression such as "believed," "considered," "estimated to be," "thought to be," and

so forth, as exemplified by the third of the foregoing examples.

The term *uncertainty* may sometimes be used effectively to achieve a conciseness of expression otherwise difficult or impossible to attain. Thus, one might make a statement such as:

> The uncertainties in the above values are not more than ± 0.5°C in the range 0°C to 1100°C, and then increase to ± 2°C at 1450°C,

or

> The uncertainty in this value does not exceed . . . excluding (or, including) the uncertainty of . . . in the value . . . adopted for the (reference standard involved).

A statement giving numerical limits of uncertainty as in the above should be followed by a brief discussion telling how the limits were derived.

Finally, the following forms of quasi-absolute statements are considered poor practice, and are to be avoided:

> The accuracy of . . . is 5 percent.
> The accuracy of . . . is ± 2 percent.

These are presumably intended to mean that the result concerned is not inaccurate, that is, not in error, by more than 5 percent or 2 percent, respectively, but they explicitly state the opposite.

Neither Systematic Error Nor Imprecision Negligible

When neither the imprecision nor the systematic error of a result are negligible, then the following rules are recommended:

1) A reported result should be qualified by a quasi-absolute type of statement that places bounds on its systematic error, and a separate statement of its standard error or its probable error, or of an upper bound thereto, whenever a reliable determination of such value or bound is available. Otherwise a computed value of the standard error, or, probable error, so designated, should be given together with a statement of the number of degrees of freedom on which it is based.

2) The bounds to its systematic error and the measure of its imprecision should be stated to no more than two significant figures.

3) The reported result itself should be stated at most to the last place affected by the finer of the two qualifying statements (unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that it may possess for certain particular uses).

4) The qualification of a reported result with respect to its imprecision and systematic error should be given in sentence form, except when results of different precision or with different bounds to their systematic errors are presented in tabular arrangement. If it is necessary or desirable to indicate their respective imprecisions or bounds to their respective systematic errors, such information may be given in a parallel column or columns, with appropriate identification.

Here, and in the next section, the term *standard error* is to be understood as signifying the standard deviation of the reported value itself, not as signifying the standard deviation of the single determination (unless, of course, the reported value is simply the result of a single determination).

The above recommendations should not be construed to exclude the presentation of a quasi-abolute type of statement placing bounds on the inaccuracy, that is, on the overall uncertainty, of a reported value, provided that separate statements of its imprecision and its possible systematic error are included also. To be in good taste, the bounds indicating the overall uncertainty should not be numerically less than the corresponding bounds placed on the systematic error outwardly increased by at least three times the standard error. The fourth of the following examples of good practice is an instance at point:

> The standard errors of these values do not exceed 0.000004 inch, and their systematic errors are not in excess of 0.00002 inch.

> The standard errors of these values are less than (*x* units), and their systematic errors are thought to be less than ± (*y* units). No additional uncertainty is assigned for the conversion to the chemical scale since the adopted conversion factor is taken as 1.000275 exactly.

> . . . with a standard error of (*x* units), and a systematic error of not more than ± (*y* units).

> . . . with an overall uncertainty of ± 3 percent based on a standard error of 0.5 percent and an allowance of ± 1.5 percent for systematic error.

When a reliably established value for the relevant standard error is available, and the dispersion of the present measurements is in keeping with this experience, then this canonical value of the standard error should be used (5). If such experience indicates that the standard error is subject to fluctuations greater than the intrinsic variation of such a measure, then an appropriate upper bound should be given, for example, as in the first two of the above examples, or by changing "a standard error . . ." in the third and fourth examples to "an upper bound to the standard error . . ."

When there is insufficient recent experience with the measurement processes involved, an estimate of the standard error must of necessity be computed by recognized statistical procedures from the same measurements as the reported value itself. It is essential that such computations be carried out according to an agreed-upon standard procedure, and the results thereof presented in sufficient detail to enable the reader to form his own judgment, and make his own allowances for their inherent uncertainties. To avoid possible misunderstanding, in such cases, first, the term *computed standard error* should be used; second, the estimate of the standard error employed should be that obtained from

$$\text{estimate of standard error} = \left(\frac{\text{sum of squared residuals}}{n\nu}\right)^{1/2}$$

where n is the (effective) number of completely independent determinations of which a is the arithmetic mean (or other appropriate least-squares adjusted value) and ν is the number of degrees of freedom involved in the sum of squared residuals (that is, the number of residuals minus the number of fitted constants or other independent constraints on the residuals); and third, the number of degrees of freedom should be explicitly stated. If the reported value a is the arithmetic mean, then:

$$\text{estimate of standard error} = (s^2/n)^{1/2}$$

where

$$s^2 = \sum_{i=1}^{n} (x_i - a)^2/(n-1)$$

and n is the number of completely independent determinations of which a is the arithmetic mean. For example:

> . . . which is the arithmetic mean of (*n*) independent determinations and has a standard error of . . .

> . . . with an overall uncertainty of ± 5.2 km/sec based on a standard error of 1.5 km/sec and estimated bounds of ± 0.7 km/sec on the systematic error. (The figure 5.2 is equal to 0.7 plus 3 times 1.5.)

or, if based on a computed standard error,

> The computed probable error (or, standard error) of these values is (*x* units),

based on (ν) degrees of freedom, and the systematic error is estimated to be less than $\pm$ (y units).

. . . with an overall uncertainty of $\pm$ 7 km/sec derived from bounds of $\pm$ 0.7 km/sec on the systematic error and a computed standard error of 1.5 km/sec based on 9 degrees of freedom. [The number 7 is approximately equal to $0.7 + (4.3 \times 1.5)$, where 4.3 is the value of Student's t for 9 degrees of freedom exceeded in absolute value with 0.002 probability. As $\nu \to \infty$, $t_{.002}(\nu) \to 3.090$.]

When the reported value is the result of a complex measurement process and is obtained as a function of several quantities whose standard errors have been computed, these several quantities and their standard errors should usually be reported, together with a description of the method of computation by which the standard errors were combined to provide an overall estimate of imprecision for the reported value.

Systematic Error Negligible, Imprecision Not Negligible

When the systematic error of a result is negligible but its imprecision is not, the following rules are recommended:

1) Qualification of a reported value should be limited to a statement of its standard error or of an upper bound thereto, whenever a reliable determination of such value or bound is available. Otherwise a computed value of the standard error, so designated, should be given together with a statement of the number of degrees of freedom on which it is based.

2) The standard error or upper bound thereto, should be stated to not more than two significant figures.

3) The reported result itself should be stated at most to the last place affected by the stated value or bound to its imprecision (unless it is desired to indicate and preserve such relative precision of a higher order that it may possess for certain particular uses).

4) The qualification of a reported result with respect to its imprecision should be given in sentence form, except when results of different precision are presented in tabular arrangement and it is necessary or desirable to indicate their respective imprecisions in which event such information may be given in a parallel column or columns, with appropriate identification.

5) The fact that the systematic error is negligible should be stated explicitly.

The above recommendations should not be construed to exclude the presentation of a quasi-absolute type of statement placing bounds on its possible inaccuracy, provided that a separate statement of its imprecision is included also. To be in good taste, such bounds to its inaccuracy should be numerically equal to at least three times the stated standard error. The fourth of the following examples of good practice is an instance at point.

The standard errors of these values are less than (x units).

. . . with a standard error of (x units).

. . . with a computed standard error of (x units) based on (ν) degrees of freedom.

. . . with an overall uncertainty of $\pm$ 4.5 km/sec derived from a standard error of 1.5 km/sec. (The figure 4.5 is equal to 3×1.5.)

or, if based on a computed standard error,

. . . with an overall uncertainty of $\pm$ 6.5 km/sec derived from a computed standard error of 1.5 km/sec (based on 9 degrees of freedom). (The number 6.5 is equal to 4.3×1.5, where 4.3 is the value of Student's t for 9 degrees of freedom exceeded in absolute value with 0.002 probability. As $\nu \to \infty$, $t_{.002}(\nu) \to 3.09$ . .)

The remarks with regard to a computed standard error in the preceding section apply with equal force to the last two examples above.

Conclusion

The foregoing recommendations call for fuller and sharper detail than is general in common practice. They should be regarded as minimum standards of good practice. Of course, many instances require fuller treatment than that recommended here.

Thus, in the case of determinations of the "fundamental physical constants" and other basic properties of nature, the author or authors should give a detailed account of the various components of imprecision and systematic error, and list their respective individual magnitudes in tabular form, so that (i) the state of the art will be more clearly revealed, (ii) each individual user of the final result may decide for himself which of the indicated components of imprecision or systematic error are, or are not, relevant to his use of the final result, and (iii)—most important—the final result itself or its uncertainty can be modified appropriately in the light of later advances. This is, and has long been, the practice followed in the best reports of fundamental studies, but current efforts to prepare critically evaluated standard reference data have revealed that far too great a fraction of the data in the scientific literature "cannot be critically evaluated because the minimum of essential information is not present" (*6*).

References and Notes

1. The true value defined conceptually by an exemplar measurement process, or the target value intended in a practical measurement process.
2. The standard error is the standard deviation of the probability distribution of estimates (that is, reported values) of the quantity that is being measured. See M. G. Kendall and W. R. Buckland, *A Dictionary of Statistical Terms* (Hafner, New York 1957).
3. For a comprehensive discussion on precision and accuracy, and a selected bibliography of 80 references, see C. Eisenhart, "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems," *J. Res. Nat. Bur. Std.* **67C**, No. 2, 161–187 (1963). (Reprints are available upon request.)
4. The essential elements of the present recommendations first appeared in a 1955 National Bureau of Standards task group report prepared principally by Malcolm W. Jensen (Office of Weights and Measures), Leroy W. Tilton (Optics and Metrology Division), and Churchill Eisenhart (Applied Mathematics Division), which was based for the most part on detailed recommendations developed some years earlier by Dr. Tilton for the internal guidance of the Optics and Metrology Division. In September 1961, new introductory material was added to the recommendations of the 1955 task group; a few minor changes were made in the illustrative examples, and the resulting revised version was circulated as a working paper of the Subcommittee on Accuracy Statements of the NBS Testing and Calibration Committee. This 1961 version was incorporated without essential change as chapter 23, "Expression of the Uncertainties of Final Results," of NBS Handbook 91, *Experimental Statistics* (U.S. Government Printing Office, Washington, 1963), reprinted with corrections in 1966. (This handbook brought together in a single volume the material on experimental statistics prepared at the National Bureau of Standards for the U.S. Army *Ordnance Engineering Design Handbook*, and printed in 1962 for limited distribution as U.S. Army Ordnance Corps Pamphlets ORDP 20-110 through 20-114. Subsequently, when these five pamphlets became parts of the *AMC Engineering Design Handbook*, they were designated Army Materiel Command Pamphlets AMCP 706-110 through 706-114.)

   In the present version, the content of chapter 23 has been rearranged and, in order to be more appropriate to calibration work, more explicit consideration has been given to the case where the value of the standard deviation σ of the measurement process involved has been well established by recent past experience. A terse summary of the principal recommendations of the present paper in the form of a text figure (Fig. 1) is contained in H. H. Ku, "Expressions of Imprecision, Systematic Error, and Uncertainty Associated with a Reported Value," to be published in *Measurements and Data*. The earlier versions were addressed primarily to the case of isolated experiments or tests, where the relevant value of σ is usually unknown in advance, and the statistical uncertainty of the final results must therefore be expressed entirely in terms of quantities derived from the data of the experiment itself.
5. The control chart is an invaluable tool in providing justification for the use of a canonical value of the standard error. See, for example, ASTM Manual on *Quality Control of Materials* (American Society for Testing and Materials, Philadephia, 1951).
6. L. M. Branscomb, "The misinformation explosion: Is the literature worth reviewing?," a talk presented to the Philosophical Society of Washington, 17 November 1967, and to be published in *Scientific Research*.

EXPRESSIONS OF IMPRECISION, SYSTEMATIC ERROR, AND UNCERTAINTY ASSOCIATED WITH A REPORTED VALUE

HARRY H. KU, National Bureau of Standards

Reprinted with corrections, November 1968.

The work of a calibration laboratory may be thought of as a sequence of operations that result in the collection, storage, and transmittal of information. In making a statement of uncertainty of the result of calibration, the calibration laboratory transmits information to its clients on the particular item calibrated.

It is logical, then, to require the transmitted information to be meaningful and unambiguous, and to contain all the relevant information in the possession of the laboratory. *The information content of the statement of uncertainty determines, to a large extent, the worth of the calibrated value.*

A common deficiency in many statements of uncertainty is that they do not convey all the information a calibration laboratory has to offer, information acquired through much ingenuity and hard work. This deficiency usually originates in two ways:

1. Loss of information through oversimplification, and

2. loss of information through the inability of the laboratory to take into account information accumulated from its past experience.

With the increasingly stringent demands for improved precision and accuracy of calibration work, calibration laboratories as a whole just cannot afford such luxury.

Traceability to the national standards, accuracy ratios, and class tolerance requirements are simplified concepts that aim to achieve different degrees of accuracy requirements. These concepts and the resulting statements are useful on certain occasions, but fail whenever the demand is exacting. The general practice of obliterating all the identifiable components of uncertainty, by combining them into an overall uncertainty, just for the sake of simplicity, is another case in point. After all, if the calibration laboratory reports all the pertinent information in separate components, the user can always combine them or use them individually, as he sees fit. On the other hand, if the user is given only one number, he can never disentangle this number into its various components. Since the information buried under these oversimplified statements is available, and may well be useful to sophisticated customers, such practices result in substantial waste of effort and resources.

In calibrating an item by repeating the same calibration procedure, the calibration laboratory gains increments of information about its calibration *system*. These increments of information are quantified and accumulated for the benefit of the calibration laboratory. If the precision of the calibration process remains unchanged, the statistical measure of dispersion (s) – i.e., the standard deviations computed from these sets of data – can be pooled together, weighted by their respective degrees of freedom. When many such increments of information are combined, an accepted or canonical value of standard deviation (σ) is established. This established (canonical) value of standard deviation characterizes the precision of the calibration process, and is treasured information in any calibration laboratory.

HARRY H. KU has been a mathematical statistician in the Statistical Engineering Laboratory, National Bureau of Standards, Washington, since 1959. He received a MSCE from Purdue University in 1941, and a Ph.D. in mathematical statistics from George Washington University in 1968. He is a member of the American Statistical Association and the Institute of Mathematical Statistics. Since 1964 he has been a member of Advisory Group 5.3, Application of Statistical Methods, of the American Standards Association's Committee B-89 on Dimensional Metrology. In 1965 he became a consulting member of subcommittees V (Analysis of Data) and VI (Statistical Nomenclature and Definitions) of ASTM Committee E-11 on Statistical Methods.

Hence, the canonical value of standard deviation is the quantification of information accumulated from past experiences of the calibration laboratory, and is an essential element of the statement of uncertainty. The standard deviation (s) computed from the current calibration is used to check the precision of current work, and to add to the pool of information on the process, but certainly does not represent all the information available in the possession of an established calibration laboratory. Only by passing its accumulated information to the users is the calibration laboratory performing a complete service.

STATEMENT OF UNCERTAINTY

Hence, in the preparation of a statement of uncertainty, it is helpful to bear in mind that:

1. The derivation of a statement of uncertainty has as its foundation the work done in the laboratory, and is based on information accumulated from past experience, and

2. In general, information is lost through oversimplification, and demands for improved precision and accuracy cannot be met with simplified statements of uncertainty.

Unless a statement of uncertainty is well formulated and supported, it is difficult to say what is meant by the statement, a difficulty frequently encountered. Since the evaluation of uncertainty is part and parcel of the high standard of work of a calibration laboratory, the statement of uncertainty deserves all the attention required to make the statement both realistic and useful. To this end, Tables 1, 2 and 3 give terms and expressions compiled as a ready reference for those who are searching for some appropriate format or wording, to carry out the thoughts expressed. They summarize the recommended practices on expression of uncertainties as given in Chapter 23 of NBS Handbook 91. A revised version of this chapter with the title "Expression of Uncertainties of Final Results" by Churchill Eisenhart may be found in *Science*, 160, June 14, 1968. Figure 1 gives a condensed summary of this material. Tables 1, 2 and 3 give details on the following:

IMPRECISION
- Standard deviation
- Standard error
- Confidence interval
- Probable error
- Mean deviation
- Arithmetic mean
- Weighted mean
- Fitted equation

SYSTEMATIC ERROR
- Uncertainty in constants
- Uncertainty in calibrated values
- Bias in computation

UNCERTAINTY
- Bounds to inaccuracy

TABLE 1 – IMPRECISION STATEMENTS

Value reported	Index or Measure of Error	Remarks
Precision of a measurement (calibration) process	(a). Standard deviation (σ) of a single determination (observation)	σ (or s with the associated degrees of freedom[1]) is of main interest as an index of precision of the measurement process. If the average of n such measurements is also reported, see (b) below.
Arithmetic mean ($\bar{x}_n$) of n numbers	(b). Standard error ($\sigma/\sqrt{n}$) of the reported value	$\bar{x}_n$ is of main interest; the number n is also essential information; σ assumed known.[1]
	(c). 2 sigma limits (d). 3 sigma limits	Commonly used bounds of imprecisions; usually used when σ known, or when n large.
	(e). Confidence interval (indicate one- or two-sided)	Data points assumed to be normally distributed; report confidence coefficient (level) $100(1-\alpha)\%$. [2]
	(f). Half-width of confidence interval (or confidence limits)	Same as (e) above; for symmetrical two-sided intervals; an index to bounds of imprecision. [2]
	(g). Probable error of the reported value	Probable error $= .6745\,\frac{\sigma}{\sqrt{n}}$ for normally distributed data points when σ known. Use of $\sigma/\sqrt{n}$ preferred. Incorrect if σ not known.
	(h). Mean deviation, or average deviation, of a measurement from the mean calculated from the sample	Limiting mean of mean deviation $= \sqrt{\frac{2}{\pi}}\,\sqrt{\frac{n-1}{n}}\,\sigma$ for normally distributed data points when σ known. Use of σ usually preferred.
	(i). Any of the above expressed in percent, or ppm of $\bar{x}_n$.	State what is being expressed in percent, eg., $(\sigma/\sqrt{n})(100/\bar{x}_n)$, $\bar{x}_n$ being a fairly constant value.
m means each computed from n measurements	(j). (b), (c), (d) and (f) above	If the measurements are of equal precision and σ unknown, use $s_p^2 = \frac{1}{m}\sum_{i=1}^{m} s_i^2$ as estimate of σ^2. The no. of degrees of freedom associated with s_p is $m(n-1)$.
	(k). Sample coefficient of variation ($v = \frac{s}{\bar{x}_n}$) or relative percent ($v \times 100$)	Appropriate when the m means cover a wide range and where the v's computed for the m sets are about the same magnitude. Give range of v's for the m sets. The means must be positive and bounded away from zero.
Weighted mean $\bar{\bar{x}} = \frac{w_1\bar{x}_1 + w_2\bar{x}_2}{w_1 + w_2}$	(l). Standard error ($\sigma_{\bar{\bar{x}}}^2$) of the weighted mean	If $w_1 = 1/\sigma_{\bar{x}_1}^2$ and $w_2 = 1/\sigma_{\bar{x}_2}^2$, then $\sigma_{\bar{\bar{x}}}^2 = \frac{1}{w_1+w_2}$. Not recommended when the σ's are not known and are estimated by s computed from small number of measurements.
An equation (theoretical or empirical) fitted to data points by the method of least squares	(m). Standard deviation computed from the deviations (residuals) of data points from the fitted curve	Report n, the number of data points, and k, the number of constants fitted, $s^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-k)$, where $\hat{y}_i$ is the value on the fitted curve for the particular x_i.[3] Value of s usually given in computer print-out.
Constants (coefficients) in the equation fitted to the data points by the method of least squares	(n). Standard errors of the coefficients based on the standard deviation computed under (m)	Standard errors usually given in computer print-out. Report n and k as above.[3].

TABLE 1 – IMPRECISION STATEMENTS – (Continued)

Value reported	Index or Measure of Error	Remarks
A predicted point on the curve $\hat{y}$ for a particular x_o	(o). Standard error ($s_{\hat{y}}$) of the predicted point	For the straight line case, the computer print-out gives the variance-covariance matrix $\begin{pmatrix} s_{11} & \\ s_{12} & s_{22} \end{pmatrix}$. $s_{\hat{y}}^2 = s_{11} + 2s_{12}x_o + s_{22}x_o^2$.[3] Report n and k.
A predicted observed value for a particular x_o	(p). Standard error of the predicted value of y	For the straight line case, $s_y^2 = s_{\hat{y}}^2 + s^2$ where $s_{\hat{y}}^2$ and s^2 are that given in (o) and (m) respectively.[3] Report n and k
Value of function of the arithmetic means of several measured variables	(q). Standard error calculated by the use of propagation of error formulas	Appropriate when errors of measurements are small compared to the values of variables measured. Use standard error of the means of the variables in the formulas.[4] Report number of measurements from which these standard errors are computed.
Percentage or proportion (r/ n), r and n being counts	(r). Confidence limits of the true proportion P	Procedures for obtaining exact and approximate confidence limits are discussed in Chapter 7, NBS Handbook 91. State one-sided or two-sided.

TABLE 2 – SYSTEMATIC ERROR[5] (BIAS) STATEMENTS

Value reported	Index or Measure of Error	Remarks
Numerical value resulting from a measurement process	Reasonable bounds ascribed to the value originating from: (i). systematic error reliably established	Detailed discussions of systematic errors are always helpful. Positive wording is appropriate: "... is not in error by more than ..." "... is accurate within ± ..."
	(ii). systematic error estimated from experience or by judgment	Use modifier such as "believed", "estimated", "considered", to signify the conjectural nature of the statement.
	(iii). combination of a number of elemental systematic errors	State explicitly the method of combination such as "the simple sum of the bounds" or "the square root of the sum of squares".
	(iv). uncertainty in some fundamental constant	Give reference to the value of constant used.
	(v). uncertainty in calibrated values	Ascertain the meaning of the systematic and random components of the uncertainty from the calibration laboratory so that decisions on the uses of these components can be made from the correct interpretations.
	(vi). bias in the method of computation	Correct if feasible, or give the magnitude; an example is ratio of the averages versus average of the ratios.

TABLE 3 – UNCERTAINTY STATEMENTS

Value reported	Index or Measure of error	Remarks
Numerical value resulting from a measurement process	Bounds to inaccuracy: (1). Systematic error and imprecision both negligible	Explicit expression of correctness to the last significant figure, interpreted as being accurate within $\pm 1/2$ units in the last significant figure given.
	(2). Imprecision negligible. Bounds on inaccuracy given to no more than two significant figures.	Sentence form preferred such as given under remark for (i) and (ii). Footnote needed if bounds are given in tabular form.
	(3). Systematic error negligible. Index of precision (b), (g), (h), (i), (k), or (n) stated to no more than two significant figures	State explicitly the index used and give essential information associated with the index. Qualify index calculated by the word "computed". Avoid using expressions of the form a ± b unless the meaning of b is explained fully immediately following or in footnote.
	(3'). Systematic error negligible. Bounds to imprecision (c), (d), (e), or (f) stated to no more than two significant figures.	Same as under (3).
	(4). Neither systematic error nor imprecision negligible. Two numerics indicating bounds to systematic error and index of imprecision respectively	(2) and (3) above separately stated.
	(4'). Bounds to systematic error and imprecision combined, indicating the likely inaccuracy of the value	(2) and (3') above where the two components either have been previously described, or explained immediately following (or in footnote).
	(5). Quoted from literature	State reference and give author's interpretation of the uncertainty; add remark if meaning unknown or ambiguous.

[1] If σ is not known, use the computed standard deviation s based on k measurements as an estimate of σ, where $s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (x_i - \bar{x}_k)^2$. *The number (k–1) is the degrees of freedom associated with s.*

[2] For interpretation see Chapter 1, NBS Handbook 91, **Experimental Statistics**, by M. G. Natrella, 1963.

[3] For details see Chapter 5 (straight line), and Chapter 6 (multivariate and polynomial), NBS Handbook 91.

[4] For details see "Notes on the use of propagation of error formulas", by Harry H. Ku, NBS Journal of Research, Vol. 70C, No. 4, October-December, 1966.

[5] See "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems" by Churchill Eisenhart, NBS Journal of research, Vol. 67C, No. 2, April-June, 1963, and "Systematic Errors in Physical Constants" by W. J. Youden, Physics Today 14, 1961.

FIGURE 1 – SUMMARY OF RECOMMENDATIONS ON EXPRESSIONS OF THE UNCERTAINTIES OF FINAL RESULTS

SYSTEMATIC ERROR AND IMPRECISION BOTH NEGLIGIBLE (CASE 1)

In this case, the reported result should be given correct to the number of significant figures consistent with the accuracy requirements of the situation, together with an explicit statement of its accuracy or correctness.

SYSTEMATIC ERROR NOT NEGLIGIBLE, IMPRECISION NEGLIGIBLE (CASE 2)

(a) Qualification of a reported result should be limited to a single quasi-absolute type of statement that places bounds on its inaccuracy;

(b) These bounds should be stated to no more than two significant figures;

(c) The reported result itself should be given (i.e., rounded) to the last place affected by the stated bounds, unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that the result may possess for certain particular uses;

(d) Accuracy statements should be given in sentence form in all cases, except when a number of results of different accuracies are presented, e.g., in tabular arrangement. If it is necessary or desirable to indicate the respective accuracies of a number of results, the results should be given in the form $a \pm b$ (or $a \pm \frac{b}{c}$, if necessary) with an appropriate explanatory remark (as a foot-note to the table, or incorporated in the accompanying test) to the effect that the $\pm b$, or $\pm \frac{b}{c}$, signify bounds to the errors which the a's may be subject.

(e) The fact that the imprecision is negligible should be stated explicity.

NEITHER SYSTEMATIC ERROR NOR IMPRECISION NEGLIGIBLE (CASE 3)

(a) A reported result should be qualified by: (1) a quasi-absolute type of statement that places bounds on its systematic error; and, (2) a separate statement of its standard error or its probable error or of an upper bound thereto, whenever a reliable determination of such value or bound is available – otherwise, a computed value of the standard error or probable error so designated should be given, together with a statement of a number of degrees of freedom on which it is based;

(b) The bounds to its systematic error and the measure of its imprecision should be stated to no more than two significant figures;

(c) The reported result itself should be stated, at most, to the last place affected by the finer of the two qualifying statements, unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that the result may possess for certain particular uses;

(d) The qualification of a reported result, with respect to its imprecision and systematic error, should be given in sentence form, except when results of different precision or with different bounds to their systematic errors are presented in tabular arrangement. If it is necessary or desirable to indicate their respective imprecisions or bounds to their respective systematic errors, such information may be given in a parallel column or columns, with appropriate identification.

SYSTEMATIC ERROR NEGLIGIBLE, IMPRECISION NOT NEGLIGIBLE (CASE 4)

(a) Qualification of a reported value should be limited to a statement of its standard error or of an upper bound thereto, whenever a reliable determination of such value or bound is available. Otherwise, a computed value of the standard error so designated should be given, together with a statement of the number of degrees of freedom on which it is based;

(b) The standard error, or upper bound thereto, should be stated to not more than two significant figures;

(c) The reported result itself should be stated, at most, to the last place affected by the stated value or bound to its imprecision, unless it is desired to indicate and preserve such relative precision of a higher order that the result may possess for certain particular uses;

(d) The qualification of a reported result with respect to its imprecision should be given in sentence form, except when results of different precision are presented in tabular arrangement and it is necessary or desirable to indicate their respective imprecisions, in which event such information may be given in a parallel column or columns, with appropriate identification.

(e) The fact that the systematic error is negligible should be stated explicitly.

2. Design of Experiments in Calibration

Foreword

Statistical design deals with the scheduling and the orderly arrangement of the sequence of observations in an experiment. Since each experiment is an individual undertaking, so is its design. Some basic cosiderations, however, are applicable to almost all experiments. These principles are summarized in Chapter 11 of Handbook 91, reprinted here as the first paper (2.1) in this section.

In Physical Measurements and Experiment Design (2.6), Youden highlighted the shift of emphasis from the classical designs for agricultural experimentation to that for physical experimentation. He argues that the designs should take advantage of the special features that are characteristic of the class of problems in physical sciences, and gives several examples illustrating his point.

His other three papers (2.2, 2.3, 2.5) are coauthored with scientists in various areas of the Bureau, and illustrate the need for tailoring the design to the particular experiment. A variety of other examples are also given in Statistical Design (Selected Reference D2), which is a collection of his bimonthly articles from Industrial and Engineering Chemistry.

In calibration work, it is not uncommon for different laboratories to use the same method of comparison for the same type of standards, and hence to use the same type of design. Design for the comparison of groups of standard cells are given in NBS Technical Note 430 (abstracted in 7.2). Current designs used in the comparison of mass standards are illustrated in Technical Note 288 (abstracted in 7.1). It is expected that more of this "standard type" of calibration designs for physical quantities that are routinely measured in Standards Laboratories will be published in the form of Technical Notes. One such publication in preparation is that for the series of mass standards.

CHAPTER 11*

GENERAL CONSIDERATIONS IN PLANNING EXPERIMENTS

Mary G. Natrella

11-1 THE NATURE OF EXPERIMENTATION

An experiment has been defined, in the most general sense, as "a considered course of action aimed at answering one or more carefully framed questions." Observational programs in the natural sciences and sample surveys in the social sciences are clearly included in this general definition. In ordnance engineering, however, we are concerned with a more restricted kind of experiment in which the experimenter *does something* to at least some of the things under study and then *observes the effect of his action.*

The things under study which are being deliberately varied in a controlled fashion may be called the *factors.* These factors may be quantitative factors such as temperature which can be varied along a continuous scale (at least for practical purposes the scale may be called continuous) or they may be qualitative factors (such as different machines, different operators, different composition of charge, etc.). The use of the proper *experimental pattern* aids in the evaluation of the factors. See Paragraph 11-2.

In addition to the factors, which are varied in a controlled fashion, the experimenter may be aware of certain background variables which might affect the outcome of the experiment. For one reason or another, these background variables will not be or cannot be included as factors in the experiment, but it is often possible to plan the experiment so that:

(1) possible effects due to background variables do not affect information obtained about the factors of primary interest; and,

(2) some information about the effects of the background variables can be obtained. See Paragraph 11-3.

In addition, there may be variables of which the experimenter is unaware which have an effect on the outcome of the experiment. The effects of these variables may be given an opportunity to "balance out" by the introduction of *randomization* into the experimental pattern. See Paragraph 11-4.

Many books have been written on the general principles of experimentation, and the book by Wilson[1] is especially recommended. There are certain characteristics an experiment obviously must have in order to accomplish anything at all. We might call these *requisites of a good experiment,* and we give as a partial listing of requisites:

(1) There must be a clearly defined objective.

(2) As far as possible, the effects of the factors should not be obscured by other variables.

(3) As far as possible, the results should not be influenced by conscious or unconscious bias in the experiment or on the part of the experimenter.

* NBS Handbook 91, 1966.

(4) The experiment should provide some measure of precision.*

(5) The experiment must have sufficient precision to accomplish its purpose.

* This requisite can be relaxed in some situations, i.e., when there is a well-known history of the measurement process, and consequently good *a priori* estimates of precision.

To aid in achieving these requisites, statistical design of experiments can provide some *tools for sound experimentation*, which are listed in Table 11-1.

The tools given include: *experimental pattern, planned grouping, randomization*, and *replication*. Their functions in experimentation are shown in Table 11-1, and are amplified in Paragraphs 11-2 through 11-5.

TABLE 11-1. SOME REQUISITES AND TOOLS FOR SOUND EXPERIMENTATION

Requisites	Tools
1. The experiment should have carefully defined objectives.	1. The definition of objectives requires all of the specialized subject-matter knowledge of the experimenter, and results in such things as: (a) Choice of factors, including their range; (b) Choice of experimental materials, procedure, and equipment; (c) Knowledge of what the results are applicable to.
2. As far as possible, effects of factors should not be obscured by other variables.	2. The use of an appropriate EXPERIMENTAL PATTERN** (see Par. 11-2) helps to free the comparisons of interest from the effects of uncontrolled variables, and simplifies the analysis of the results.
3. As far as possible, the experiment should be free from bias (conscious or unconscious).	3. Some variables may be taken into account by PLANNED GROUPING (see Par. 11-3). For variables not so taken care of, use RANDOMIZATION (Par. 11-4). The use of REPLICATION aids RANDOMIZATION to do a better job.
4. Experiment should provide a measure of precision (experimental error).*	4. REPLICATION (Par. 11-5) provides the measure of precision; RANDOMIZATION assures validity of the measure of precision.
5. Precision of experiment should be sufficient to meet objectives set forth in requisite 1.	5. Greater precision may be achieved by: Refinements of technique EXPERIMENTAL PATTERN (including PLANNED GROUPING) REPLICATION.

* Except where there is a well-known history of the measurement process.

** Capitalized words are discussed in the following paragraphs.

11-2 EXPERIMENTAL PATTERN

The term *experimental pattern* is a broad one by which we mean the planned schedule of taking the measurements. A particular pattern may or may not include the succeeding three tools (*planned grouping*, *randomization*, and *replication*). Each of these three tools can improve the experimental pattern in particular situations. The proper pattern for the experiment will aid in control of bias and in measurement of precision, will simplify the requisite calculations of the analysis, and will permit clear estimation of the effects of the factors.

A common experimental pattern is the so-called factorial design experiment, wherein we control several factors and investigate their effects at each of two or more levels. If two levels of each factor are involved, the experimental plan consists of taking an observation at each of the 2^n possible combinations. The factorial design, with examples, is discussed in greater detail in Chapter 12.

11-3 PLANNED GROUPING

An important class of experimental patterns is characterized by *planned grouping*. This class is often called *block designs*. The use of planned grouping (blocking) arose in comparative experiments in agricultural research, in recognition of the fact that plots that were close together in a field were usually more alike than plots that were far apart. In industrial and engineering research, the tool of planned grouping can be used to take advantage of naturally homogeneous groupings in materials, machines, time, etc., and so to take account of "background variables" which are not directly "factors" in the experiment.

Suppose we are required to compare the effect of five different treatments of a plastic material. Plastic properties vary considerably within a given sheet. To get a good comparision of the five treatment effects, we should divide the plastic sheet into more or less homogeneous areas, and subdivide each area into five parts. The five treatments could then be allocated to the five parts of a given area. Each set of five parts may be termed a block. In this case, had we had four or six treatments, we could as well have had blocks of four or six units. This is not always the case — the naturally homogeneous area (block) may not be large enough to accommodate all the treatments of interest.

If we are interested in the wearing qualities of automobile tires, the natural block is a block of four, the four wheels of an automobile. Each automobile may travel over different terrain or have different drivers. However, the four tires on any given automobile will undergo much the same conditions, particularly if they are rotated frequently.

In testing different types of plastic soles for shoes, the natural block consists of two units, the two feet of an individual.

The block may consist of observations taken at nearly the same time or place. If a machine can test four items at one time, then each run may be regarded as a block of four units, each item being a unit.

Statisticians have developed a variety of especially advantageous configurations of *block designs*, named and classified by their structure into randomized blocks, Latin squares, incomplete blocks, lattices, etc., with a number of subcategories of each. Some of these block designs are discussed in detail in Chapter 13.

11-4 RANDOMIZATION

Randomization is necessary to accomplish Requisites 3 and 4 in Table 11-1. In order to eliminate bias from the experiment (Requisite 3), experimental variables which are not specifically controlled as factors, or "blocked out" by planned grouping, should be randomized — e.g., the allocations of specimens to treatments or methods should be made by some mechanical method of randomization.

Randomization also assures valid estimates of experimental error (Requisite 4), and makes possible the application of statistical tests of significance and the construction of confidence intervals.

There are many famous examples of experiments where failure to randomize at a crucial stage led to completely misleading results. As always, however, the coin has another side; the beneficial effects of randomization are obtained in the long run, and not in a single isolated experiment. Randomization may be thought of as insurance, and, like insurance, may sometimes be too expensive. If a variable is thought unlikely to have an effect, and if it is very difficult to randomize with respect to the variable, we may choose not to randomize.

In general, we should try to think of all variables that could possibly affect the results, select as factors as many variables as can reasonably be studied, and use planned grouping where possible. Ideally, then, we randomize with respect to everything else — but it must be recognized that the ideal cannot always be realized in practice.

The word *randomization* has been used rather than *randomness* to emphasize the fact that experimental material rarely, if ever, has a random distribution in itself, that we are never really safe in assuming that it has, and that consequently randomness has to be assured by formal or mechanical randomization.

11-5 REPLICATION

In order to evaluate the effects of factors, a measure of precision (experimental error) must be available. In some kinds of experiments, notably in biological or agricultural research, this measure must be obtained from the experiment itself, since no other source would provide an appropriate measure. In some industrial and engineering experimentation, however, records may be available on a relatively stable measurement process, and this data may provide an appropriate measure. Where the measure of precision must be obtained from the experiment itself, *replication* provides the measure. In addition to providing the measure of precision, replication provides an opportunity for the effects of uncontrolled factors to balance out, and thus aids randomization as a bias-decreasing tool. (In successive replications, the randomization features must be independent.) Replication will also help to spot gross errors in the measurements.

11-6 THE LANGUAGE OF EXPERIMENTAL DESIGN

In discussing applications of statistical design of experiments in the field of physical sciences and engineering, we are extremely handicapped by the classical language of experimental design. The early developments and applications were in the field of agriculture, where the terms used in describing the designs had real physical meaning. The *experimental area* was an area — a piece of ground. A *block* was a smaller piece of ground, small enough to be fairly uniform in soil and topography, and thus was expected to give results within a block that would be more alike than those from different blocks. A *plot* was an even smaller piece of ground, the basic unit of the design. As a unit, the plot was planted, fertilized, and harvested, and it could be *split* just by drawing a line. A *treatment* was actually a treatment (e.g., an application of fertilizer) and a *treatment combination* was a combination of treatments. A *yield* was a yield, a quantity harvested and weighed or measured.

Unfortunately for our purposes, these are the terms commonly used. Since there is no particular future in inventing a new descriptive language for a single book, we must use these terms, and we must ask the engineer or scientist to stretch his imagination to make the terms fit his experimental situation.

Experimental area can be thought of as the scope of the planned experiment. For us, a *block* can be a group of results from a particular operator, or from a particular machine, or on a particular day — any planned natural grouping which should serve to make results from one block more alike than results from different blocks. For us, a *treatment* is the factor being investigated (material, environmental condition, etc.) in a single factor experiment. In factorial experiments (where several variables are being investigated at the same time) we speak of a *treatment combination* and we mean the prescribed levels of the factors to be applied to an experimental unit. For us, a *yield* is a measured result and, happily enough, in chemistry it will sometimes be a yield.

Many good books on experimental design are available. See the following list of References and Recommended Textbooks.

REFERENCES

1. E. B. Wilson, Jr., *An Introduction to Scientific Research*, McGraw-Hill Book Co., Inc., New York, N.Y., 1952.

SOME RECOMMENDED TEXTBOOKS

R. L. Anderson and T. A Bancroft, *Statistical Theory in Research*, McGraw-Hill Book Co., Inc., New York, N.Y., 1952.

V. Chew (ed.), *Experimental Designs in Industry*, John Wiley and Sons, Inc., New York, N.Y., 1958.

W. G. Cochran and G. M. Cox, *Experimental Designs* (2d edition), John Wiley and Sons, Inc., New York, N.Y., 1957.

D. R. Cox, *Planning of Experiments*, John Wiley and Sons, Inc., New York, N.Y., 1958.

O. L. Davies (ed.), *The Design and Analysis of Industrial Experiments*, Oliver and Boyd, Ltd., Edinburgh, and Hafner Publishing Co., New York, N.Y., 1954.

W. T. Federer, *Experimental Design*, The Macmillan Company, New York, N.Y., 1955.

R. A. Fisher, *The Design of Experiments* (7th edition), Hafner Publishing Co., New York, N.Y., 1960.

F. A. Graybill, *An Introduction to Linear Statistical Models*, Vol. I, McGraw-Hill Book Co., Inc., New York, N.Y., 1961.

O. Kempthorne, *The Design and Analysis of Experiments*, John Wiley and Sons, Inc., New York, N.Y., 1952.

M. H. Quenouille, *The Design and Analysis of Experiment*, Hafner Publishing Co., New York, N.Y., 1953.

H. Scheffé, *The Analysis of Variance*, John Wiley and Sons, Inc., New York, N.Y., 1959.

W. J. Youden, *Statistical Methods for Chemists*, John Wiley and Sons, Inc., New York, N.Y., 1951.

The problem posed in the introduction can now be resolved in many ways. If the 36 quantities are divided into two groups of 18 each, 324 pairs will be formed. At the other extreme is the division into 1 and 35, which results in only 35 pairs.

3. Application to Thermometer Calibration

The authors asked the Thermometry Section of the National Bureau of Standards to intercompare eight thermometers, using the two-group arrangement. The usual practice of the section is to read the thermometers in sequence in a bath with slowly rising temperature and then to read them in reverse order. This device effectively compensates for changes in the bath temperature, provided that the temperature changes at a constant rate. *The effectiveness of the two-group arrangement, however, does not depend on a constant rate of change in temperature.*

The thermometers were partly immersed in a bath of distilled water, and were read through a telescope mounted a short distance away. The temperature of the bath was at approximately 40° C at the start of the readings, but rose gradually throughout the experiment. There were short pauses of irregular length between pairs of readings.

The eight thermometers were divided into 2 groups of 4 each, containing thermometers 1, 2, 3, and 4 and 5, 6, 7, and 8, respectively. The readings are given in table 1 in the order in which they were obtained.[2]

The computations can be simplified by subtracting some convenient number from each observation. Accordingly, all subsequent calculations are based on the observations in table 1 after subtracting 40 from each of them.

[2] The thermometers were randomized within the pairs and the pairs within the runs.

The mathematical model underlying the statistical analysis is based on the following considerations. Let M be a reference temperature in the range of temperatures of the bath during the experiment. At the time of measurement of the jth pair of thermometers, the temperature of the bath will be $M+p_j$, where p_j is defined by this condition.

Next, suppose that the ith thermometer belongs to the jth pair, and let x_{ij} denote the observed temperature for this thermometer when the jth pair is read. Then the difference between the observed temperature x_{ij} and the true bath temperature $M+p_j$ will consist of two parts: a systematic error t_i, peculiar to the ith thermometer, and a random reading error e_{ij}, i. e.,

$$x_{ij}-(M+p_j)=t_i+e_{ij}$$

or

$$x_{ij}=M+t_i+p_j+e_{ij}.$$

By imposing the restrictions $\sum_{i=1}^{p} t_i=\sum_{j=1}^{mn} p_j=0$, M is uniquely defined.

The constants M, t_i, and p_j and the error e_{ij} are unknown but can be estimated from the data. It is assumed that the errors associated with different readings are independent and come from the same population of errors. This population is assumed to have mean zero and standard deviation σ, which may or may not be known.

The following calculations will show how to estimate the constants and the standard deviation.[3] Estimates of the t's are of especial interest, since they may be used to calibrate a new thermometer in terms of a standard. Estimates are denoted by

[3] Derivations of formulas are given in the appendix.

TABLE 1. *Temperature readings in order of time*

Run 1			Run 2			Run 3			Run 4		
Pair	Thermometer	Reading	Pair	Thermometer	Reading	Pair	Thermometer	Reading	Pair	Thermometer	Reading
		°C			°C			°C			°C
1	1	40.00	5	3	40.18	9	2	40.23	13	6	40.26
	7	39.99		8	40.18		6	40.22		3	40.28
2	5	40.08	6	7	40.07	10	8	40.24	14	7	40.15
	3	40.13		2	40.19		4	40.15		4	40.20
3	8	40.15	7	1	40.10	11	7	40.12	15	5	40.27
	2	40.17		6	40.18		3	40.20		2	40.30
4	6	40.13	8	5	40.17	12	5	40.23	16	1	40.21
	4	40.05		4	40.13		1	40.16		8	40.31

carets. For example, $\hat{t}_i$ is the estimate of t_i.

To analyze the coded data it is convenient to compute an auxiliary quantity, D, for each thermometer. Thus D_i, the D for the ith thermometer, is computed as follows. For each pair that contains the ith thermometer the difference between the reading for the ith thermometer and the reading for the other thermometer of the pair is computed. The sum of these differences is D_i. For example,

$$D_1=[0-(-.01)]+(.10-.18)+(.16-.23)+(.21-.31)=-.24.$$

Let the group that contains m thermometers be called group 1, and the group that contains n thermometers be called group 2. Let the sum of the D's for the thermometers in group 1 be denoted by S_1, and in group 2 by S_2. Then the D's may be used to estimate the correction for the ith thermometer by the following formulas: [4]

$$\hat{t}_i=(vD_i-S_1)/vn$$

if i is in group 1, and

$$\hat{t}_i=(vD_i-S_2)/vm$$

if i is in group 2. For example, for the first thermometer

$$\hat{t}_1=(8D_1-S_1)/32=(-1.92+.07)/32=-.05781.$$

If σ is unknown from past experience, it may be calculated from the data. This calculation is made quite simply by working with the differences between the readings within a pair. Let the difference without regard to sign for the jth pair be designated by d_j. Then σ is estimated from the formula [5]

$$2(mn-m-n+1)\hat{\sigma}^2=\sum_{j=1}^{mn}d_j^2-\sum_{i=1}^{v}\hat{t}_iD_i.$$

The computations may be systematized by use of table 2, in which the estimates of the t's and $\sum_{i=1}^{v}\hat{t}_iD_i$ are found.

[4] It sometimes happens that the temperatures or other quantities are not observed directly, but instead the differences between the quantities in the pairs are recorded. Although in this case M and the p's cannot be estimated, the t's still are estimable by these formulas.

[5] When just the differences are observed, it is convenient to do the analysis in terms of the standard deviation of the differences, which may conveniently be denoted by σ_d. This formula and others below apply in this case, too, provided σ is replaced by $\sigma_d/2$.

Table 2. *Calculation of the thermometer effects*

		Group 2 (n) thermometer				Calculations					
		5	6	7	8	Σ	D	$8D$	$32\hat{t}$	$\hat{t}$	$D\hat{t}$
Group 1 (m) thermometer	1	.23 .16	.18 .10	-.01 0	.31 .21	.47 .71	-.24	-1.92	-1.85	-.05781	.01388
	2	.27 .30	.22 .23	.07 .19	.15 .17	.89 .71	.18	1.44	1.51	.04719	.00849
	3	.08 .13	.26 .28	.12 .20	.18 .18	.79 .54	.15	1.20	1.27	.03969	.00595
	4	.17 .13	.13 .05	.15 .20	.24 .15	.53 .59	-.16	-1.28	-1.21	-.03781	.00605
Calculations	Σ	.75 .72	.79 .66	.33 .59	.68 .71	5.43 5.43					
	D	.03	.13	-.26	.17		-.07 .07				
	$8D$	.24	1.04	-2.08	1.36			-.56 .56			
	$32\hat{t}$	.17	.97	-2.15	1.29				-.28 .28		
	$\hat{t}$	.00531	.03031	-.06719	.04031					-.00874 .00874	
	$D\hat{t}$	.00016	.00394	.01747	.00685						.03437 .02842

The coded readings are entered in the upper left-hand part of the table, where every cell corresponds to some pair. For example, the first pair is put into the cell in row 1 and column 7, with the reading for thermometer 1 recorded in the upper right-hand corner and for thermometer 7 in the lower left-hand corner. By so recording the readings, each row and column is divided into subrows and subcolumns.

The remaining rows and columns are for calculations, which it is believed are self-evident. In general, row $8D$ is replaced by vD and $32\hat{t}$ by $(vm)\hat{t}$. Likewise, column $8D$ is replaced by vD and $32\hat{t}$ by $(vn)\hat{t}$. Several checks are available: (1) the sum of the entries in row Σ must equal the sum of the entries in column Σ, and (2) the sums of the other corresponding rows and columns, except the last, must be of different sign but of the same absolute value. In the table these quantities appear along the diagonal.

The standard deviation is estimated from the formula given above. The differences d may easily be calculated from table 1, and $\sum_{i=1}^{8} t_i D_i$ from table 2. The differences and calculations on them are given in table 3.

TABLE 3. *Calculation of the standard deviation*

Pair (j)	d_j	d_j^2	Pair (j)	d_j	d_j^2
1	0.01	0.0001	9	0.01	0.0001
2	.05	.0025	10	.09	.0081
3	.02	.0004	11	.08	.0064
4	.08	.0064	12	.07	.0049
5	.00	.0000	13	.02	.0004
6	.12	.0144	14	.05	.0025
7	.08	.0064	15	.03	.0009
8	.04	.0016	16	.10	.0100

$\sum_{j=1}^{16} d_j^2 = .0651$, $\sum_{i=1}^{8} \hat{t}_i D_i = .03437 + .02842 = .0628$,

$189\hat{\sigma}^2 = .0651 - .0628 = .0023$, $\hat{\sigma} = .0114$.

Two thermometers can be compared by finding the difference between their estimated effects. To judge the significance of such a difference, it is desirable to know the standard deviation of the difference. If i and i' both are in group 1, then the square of the standard deviation of the difference is

$$\sigma^2_{\hat{t}_i - \hat{t}_{i'}} = 4\sigma^2/n;$$

if both are in group 2, then

$$\sigma^2_{\hat{t}_i - \hat{t}_{i'}} = 4\sigma^2/m;$$

and if i is in group 1 but i' is in group 2, then

$$\sigma^2_{\hat{t}_i - \hat{t}_{i'}} = 2(v-1)\sigma^2/mn;$$

If σ^2 is not known, then its estimate is used.

As an example, consider thermometers 1 and 2. The appropriate formula is the first one above, so that

$$\hat{\sigma}^2_{\hat{t}_1 - \hat{t}_2} = .0114.$$

Just as it has been possible to intercompare the thermometers even though in some cases a particular pair of thermometers were never at the same temperature, so also it is possible to determine the relative temperatures of the bath when each of the mn pairs of thermometers were read even though the temperatures were read with different thermometers with unknown corrections. It may sometimes be important to ascertain the character of the drift or changes taking place in the experimental system. In the example given, matters were arranged so that there was an approximately linear drift upward in the bath temperature. Table 4 reflects this condition, the values being computed as is indicated below.

TABEE 4. *Average temperatures of the pairs referred to 40° C*

Pair	Uncorrected average	Corrected average	Pair	Uncorrected average	Corrected average
	°C	°C		°C	°C
1	−0.005	0.058	9	0.225	0.186
2	.105	.062	10	.195	.194
3	.160	.116	11	.160	.174
4	.090	.094	12	.195	.221
5	.180	.140	13	.270	.235
6	.130	.140	14	.175	.228
7	.140	.154	15	.285	.259
8	.150	.166	16	.260	.269

The averages after correction for thermometers exhibit the upward trend much more clearly than do the crude, uncorrected averages.

The uncorrected averages for the jth pair is simply the arithmetic average of the two readings in the pair. The corrected average is the uncorrected average adjusted for the systematic errors of the thermometers that occur in the jth pair. In symbols it is $\hat{M}+\hat{p}_j$.

The estimate of M is

$$2mn\hat{M} = \sum_{i=1}^{s} \sum_{j=1}^{mn} x_{ij} + (m-n) \sum_{i=1}^{m} \hat{t}_i,$$

which, in the case at hand, reduces to

$$\hat{M} = \left(\sum_{i=1}^{8} \sum_{j=1}^{16} x_{ij} \right) / 32,$$

the grand mean of the readings. Thus $\hat{M} = 5.43/32 = .16969$. These formulas should be used with the understanding that $x_{ij}=0$ if the ith thermometer does not occur in the jth pair.

Thus far all values have been given in coded form and the adjusted thermometer readings in terms of systematic deviations from the reference temperature M. It may be of interest to estimate readings for all thermometers at temperature M. These decoded estimated readings, calculated by the formula $\hat{M}+40+\hat{t}_i$, are as follows:

Thermometer	Temperature	Thermometer	Temperature
	°C.		°C.
1	40.11	5	40.18
2	40.22	6	40.20
3	40.21	7	40.10
4	40.13	8	40.21

The estimate of p_j is obtained by a simple adjustment of the observations in the jth pair. If i and i' are the thermometers in the jth pair, then

$$2\hat{p}_j=x_{ij}+x_{i'j}-2\hat{M}-\hat{t}_i-\hat{t}_{i'}.$$

For example, for $j=2$, $i=5$, $i'=3$, and

$$2\hat{p}_2=.08+.13-2(.16969)-.00531-.03969,$$

so that $\hat{p}_2=-.08719$.

It now is possible to exhibit the decomposition of x_{52} into its parts. Thus

$$x_{52}=\hat{M}+\hat{t}_5+\hat{p}_2+\hat{e}_{52}$$

$$.08=.16969+.00531+(-.08719)+(-.00781).$$

It is interesting to note that the estimated error in this particular reading is of about the same magnitude as $\hat{\sigma}$.

The fundamental importance of the arrangement is that it makes it possible to intercompare the thermometers and to limit the error arising from fluctuations in the bath temperature to those temperature changes that take place in the very short interval required to read two thermometers. Temperature changes from one pair to another do not contribute to the error of measurement. This technique is applicable in all cases where either the apparatus or the environment may drift or undergo unpredictable changes.

4. Appendix

4.1. Derivation of Estimates

Let the group that contains m objects be denoted by G_1, and the group that contains n objects be denoted by G_2. Then the reduced normal equations for estimating the treatment (thermometer) effects are

$$n\hat{t}_i-\sum_{u=m+1}^{p}\hat{t}_u=D_i \tag{1}$$

for i in G_1 and

$$m\hat{t}_i-\sum_{u=1}^{m}\hat{t}_u=D_i \tag{2}$$

for i in G_2.

Summing over the treatments in G_1, eq (1) becomes

$$n\sum_{u=1}^{m}\hat{t}_u-m\sum_{u=m+1}^{p}\hat{t}_u=\sum_{u=1}^{m}D_u. \tag{3}$$

Then imposing the restriction

$$\sum_{u=1}^{p}t_u=0, \tag{4}$$

it is clear that

$$(n+m)\sum_{u=1}^{m}\hat{t}_u=\sum_{u=1}^{m}D_u. \tag{5}$$

Similarly, using eq (2), summing over the treatments in G_2, and applying eq (4), obtain

$$(n+m)\sum_{u=m+1}^{p}\hat{t}_u=\sum_{u=m+1}^{p}D_u. \tag{6}$$

From eq (1) and (6) it follows for i in G_1 that

$$n(n+m)\hat{t}_i=(n+m)D_i+\sum_{u=m+1}^{p}D_u$$

or since $\sum_{u=1}^{p}D_u=0$,

$$vn\hat{t}_i=vD_i-\sum_{u=1}^{m}D_u. \tag{7}$$

Similarly, for i in G_2,

$$vm\hat{t}_i=vD_i-\sum_{u=m+1}^{p}D_u. \tag{8}$$

4.2 Derivation of Variance

For random variables x and y let $V(x)$ and $Cov(x,y)$ denote, respectively, the variance of x and the covariance of x and y. Then for i and i' in G_1,

$$V(D_i)=2m\sigma^2, \quad Cov(D_i,D_{i'})=0. \tag{9}$$

From eq (7) and (9),

$$n(\hat{t}_i-\hat{t}_{i'})=(D_i-D_{i'}),$$

$$V(\hat{t}_i-\hat{t}_{i'})=4\sigma^2/n. \tag{10}$$

Similarly, for i and i' both in G_2,

$$V(\hat{t}_i-\hat{t}_{i'})=4\sigma^2/m. \tag{11}$$

For i in G_1 and i' in G_2 it is convenient to use the formula

$$V(\hat{t}_{i'}-\hat{t}_{i'})=(C_{ii}+C_{i'i'}-2C_{ii'})\sigma^2, \tag{12}$$

where C_{rs} is the element in the rth row and sth column of the inverse of the coefficient matrix of the reduced normal equations. From eq (7) and (8),

$$C_{ii}=2(v-1)/vn, \quad C_{i'i'}=2(v-1)/vm, \quad C_{ii'}=0.$$

Hence eq (12) becomes

$$V(\hat{t}_i-\hat{t}_{i'})=2(v-1)\sigma^2/mn. \tag{13}$$

4.3 Derivation of Estimate of σ

The differences d_j form a basis for the space, which consists of the error space and the space of the $\hat{t}$'s. Therefore, the sum of squares due to the d's can be partitioned into two orthogonal parts, one due to error and one due to treatments. Since the sum of squares due to treatments is $\left(\sum_{i=1}^{v}\hat{t}_iD_i\right)\Big/2$, twice the sum of squares due to error is

$$2(mn-m-n+1)\hat{\sigma}^2=\sum_{j=1}^{mn}d_j^2-\sum_{i=1}^{v}\hat{t}_iD_i.$$

WASHINGTON, September 25, 1954.

JOURNAL OF RESEARCH of the National Bureau of Standards—C. Engineering and Instrumentation
Vol. 70C, No. 2, April–June 1966

Design and Statistical Procedures for the Evaluation of an Automatic Gamma-Ray Point-Source Calibrator

S. B. Garfinkel, W. B. Mann, and W. J. Youden

Institute for Basic Standards, National Bureau of Standards, Washington, D.C.

(December 14, 1965)

A description is given of the mechanical design and operation of an automatic gamma-ray point-source calibrator.

The use of statistical design in experiments for evaluating performance factors, such as interchangeability of stations and run differences using the same data obtained in comparisons of the sources, is described in detail.

Key Words: Statistical experiment design, testing equipment, routine testing, radioactivity standardization measurements, gamma-ray point sources

1. Design and Performance of an Automatic Gamma-Ray Point-Source Calibrator

Recently, in response to a need for standards for workers in the field of gamma-ray spectrometry, a gamma-ray "kit" for point-source radioactivity standards has been developed [Hutchinson, 1960]. These sources are prepared from solutions which are standardized either by coincidence counting or, as in the case of cesium-barium-137, by measurements using the NBS calibrated $4\pi\gamma$-ionization chamber.

The sources are prepared by depositing either 0.05 or 0.1 ml of the calibrated radioactive solution onto mounts consisting of a 0.006-centimeter-thick polyester tape which is supported by an aluminum annulus (3.8 cm I.D., and 5.4 cm O.D.), as shown in figure 1. As it is desirable for all of these sources to be nominally the same strength and the same size, the solution is dispensed with an ultramicroburet [NBS Circ. 594, Mann and Seliger, 1958]. After drying, the sources are covered with another layer of the same kind of polyester tape. The sources are then intercompared with several accurately standardized sources, for the purpose of individual calibration.

For several years these calibrations were performed manually; that is, the sources were placed, one at a time, in a jig which was held in a fixed position relative to a scintillation counter, and the count rates were intercompared. As part of the program to increase the accuracy of the standards, it was decided to design and construct an automatic sample changer with the goal of attaining source intercomparisons with a precision of the order of 0.1 percent.

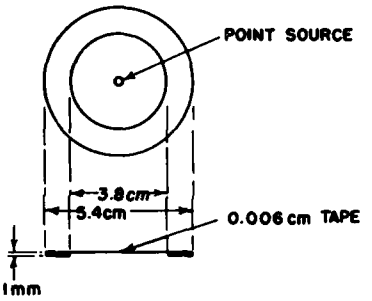


FIGURE 1. *Source mount.*

The changer is a round turn-table of 1/4-in.-thick aluminum alloy having a diameter of 24 in., with source positions spaced at 18° intervals on the circumference of a circle 20 in. in diameter (fig. 2). These positions have 1-in.-diameter holes in which rigid plastic sample carriers rest. The gamma-ray point sources are held firmly in place on top of the carriers by the pressure of phosphor-bronze springs. There are 20 indexing holes equally spaced around the table as shown in figure 2, the center of each one radially in line with the center of a sample carrier and the center of the table, and 3/8-in. in from the edge of the table. These holes, in conjunction with a solenoid-plunger pin, are used for positioning the sources above the detector.

A shaft which is affixed to the underside center of the table, rests on a steel ball bearing which lies in a conical depression inside a supporting cylinder.

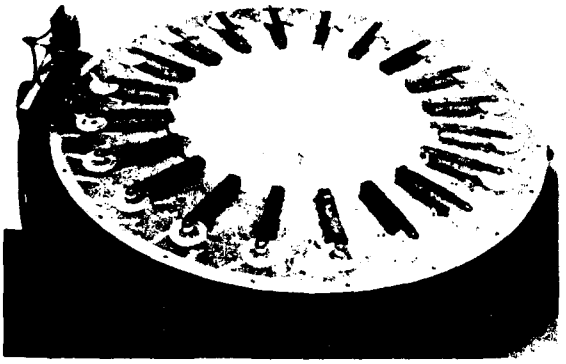


FIGURE 2. *Sample changer.*

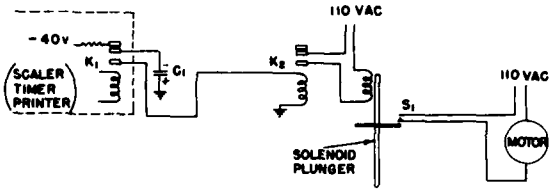


FIGURE 3. *Diagram of motor-control circuit.*

The table is rotated by a 1/100-HP motor and two gears, one of which is fixed on the motor shaft, and coupled to the other gear which is mounted on a concentric spring-loaded friction clutch on the table shaft.

The motor is turned on and off by a miniature switch (S_1), which is actuated by the plunger of a solenoid, in the following manner:

At the conclusion of a measurement, while the data are being printed out onto a paper tape, a relay, K_1 (fig. 3) in the recording system is held closed. Capacitor C_1, which had been charged up during the measurement period now discharges through the coil or relay K_2, thereby closing it for about 1.5 sec, thus energizing the solenoid. The solenoid-operated plunger is lifted from the indexing hole in the table for this brief period, and mechanically closes the miniature switch (S_1), thereby starting the motor, and the table starts to rotate. As it takes about 5 sec for the table to rotate 18°, relay K_2 opens before the next source position is reached, the solenoid is de-energized and the plunger falls back and rests on the surface of the turn table, but as this is not far enough to allow switch (S_1) to open, the motor continues to rotate the table. When the next source "arrives" into the counting position, the solenoid plunger falls into the indexing hole, thus stopping table rotation and opening the motor circuit. The purpose of the friction clutch is to allow the motor to slow down gradually after the table has stopped. The time for the sample changing is about 5.0 sec, while the printout takes 20 sec. Thus all changing operations (including the stopping of motor) stop at least 10 sec before the next measurement starts.

Originally, in order to obtain reproducible source-to-detector distance, the table was supported underneath the plunger pin by a roller bearing, and it was assumed that the combination of the spring-loaded plunger pin and the slightly loose fit of the table shaft would ensure this. However, after several series of measurements, it became apparent that sources on some positions of the table were yielding consistently erroneous values. The final design eliminated the effects of any defects in the table which would contribute to errors as a function of vertical displacement.

A lucite block with ramps at each end was affixed to the top of the lead shield, and its dimensions are such that when a source and carrier come into position, they "ride" up the ramp approximately 1.5 mm, so that the carrier is actually free of the table insofar as vertical positioning is concerned (fig. 4). The plastic sample carriers are 0.425-in. thick with a tolerance of ± 0.002 in. Thus, the source-to-detector distance is inde-

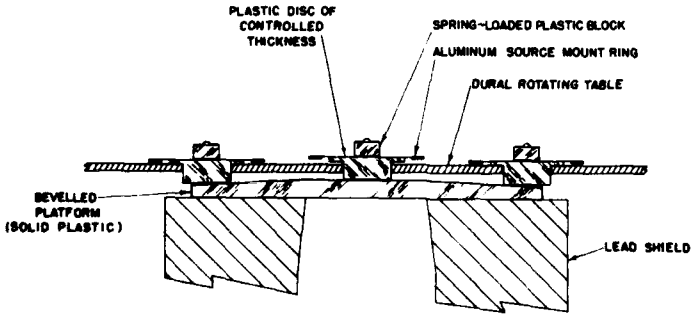


FIGURE 4. *Ramp detail.*

pendent of variations in the table thickness, and any deviations of flatness of the table. The only function of the table is to bring the sources into position above the detector, the vertical positioning being determined by the phosphor-bronze spring holding the source firmly against its carrier and the latter against the ramp. To get some idea of the reproducibility required in positioning, it should be pointed out that the source is approximately 6 in. from the detector; thus, a change in vertical position of 0.006-in. produces a change of 0.2 percent in the count rate $\left(n \propto \frac{1}{d^2}, \Delta n \propto \frac{-2\,\Delta d}{d}\right)$.

2. Description of Auxiliary Instrumentation

The gamma-ray detector consists of a 3-in. by 3-in. thallium-activated sodium iodide crystal, coupled to a 3-in. electron-multiplier phototube. The associated electronics consist of an amplifier, and gain-stabilization circuit [DeWaard, 1955], which compensates for shift in gain in either the phototube, amplifier, or high voltage supply (this latter being part of the stabilizer). The detector is situated in a lead pig, with walls 1½-in. thick (fig. 5). The aperture at the top of the shield was made small to lessen detection of unscattered gamma radiation from sources adjacent to the source being measured but large enough so that when the table rotates, the detector never "loses direct sight" of a source. Thus, the photopeak is always "present" for continuous operation of the gain-stabilizing circuit. The output from a single-channel analyzer (which is also part of the stabilizer system) whose window is set on the photopeak is fed into a commercial automatic scaler-timer-printer system. At the end of each source measurement, and after the data are printed, the scaler and timer are automatically reset, and started for the next measurement.

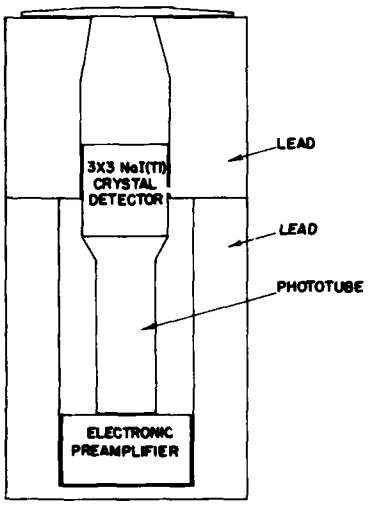


FIGURE 5. *Lead pig, showing ramp and detector assembly.*

3. Background Considerations

The activity of these sources is of the order of 5×10^4 disintegrations per second, and they are measured at a distance of about 6 in. from the 3×3-in. detector. No correction is made for the cosmic-ray background, which is of the order of 0.1 percent (or less), as variations in the background affect the ratios of nearly equal sources negligibly.

In the case of the 662-keV gamma-ray of barium-137 m, there is, for example, a relatively large background contribution (~ 4%) to the photopeak count rate arising from the detection of unscattered gamma rays from the other 19 sources. If, then, there were 19 identical sources, and the twentieth were, say, 1 percent high or low, then, the relative activity of this odd one would be in error by 0.04 percent, if, as the case is, no background corrections are made.

4. Performance

In order to assess the stability and reproducibility of the system, two experiments were performed. A cesium-137 source was put onto one of the sample carriers, and over 100 five-minute consecutive readings were taken (with no table rotation), each one consisting of some 200,000 counts. The distribution of the results fitted the expected distribution quite well.

The second investigation involved the placement of 20 sources on the table and determining (a) the relative gamma-ray emission rates of these 20 sources, as well as the bias, if any, of the 20 positions of the table. The statistical design and analyses of these experimental results are given in considerable detail. The interest centers not so much in this particular apparatus as in this type of equipment. There is increasing use of automatic equipment in the routine comparison of specimens.

5. Statistical Analysis

Industrial control laboratories and laboratories doing clinical tests are turning increasingly to mechanization of the routine operations involved in the test procedure. Sometimes these operations require the addition of reagents, mixing, and the transfer of material. The last step consists in bringing the prepared material before a testing point where a suitable device evaluates the color, *p*H, or other property of the specimen. Generally this last stage consists of a device with a number of stations which successively present their specimens to the test point.

For many tests the equivalence of the various stations is clearly satisfactory, provided only that the mechanical clearances are adequate. Should the position of the specimen, as determined by the station, be at all critical it will be necessary to demonstrate that the stations are in fact interchangeable. That is, the particular station occupied by a specimen should not contribute materially to the error in the evaluation of the specimen. Satisfactory interchangeability is desirable—the alternative being to determine suitable corrective factors for the individual stations.

There are three ways to explore experimentally the performance of the individual stations.

One procedure is to transfer the *same* specimen to every station in turn and record the reading for each station. This procedure will run into difficulty if the specimen has to be evaluated immediately, e.g., a color might fade. If the time spent at each station is fairly long, the problem of keeping the evaluating apparatus free from drift has also to be considered.

A second procedure requires the availability of as many *identical* specimens as there are stations, or of specimens which are accurately related to each other.

The above two procedures are classical and straightforward. The third procedure has the interesting feature that the stations can be evaluated while evaluating the regular sequence of specimens encountered in the work of the laboratory. The major requirement is that the specimens be stable. In brief, each specimen is evaluated at a limited number of stations, as few as three or even two stations. Each station will have been occupied by two or three or more different specimens. The values recorded will reflect the net result of the specimen plus the station characteristic. In order to obtain both the specimen values and the station corrections, there must be at least as many observations as the total of specimens and stations. Each observation can be expressed as a function of the unknown values for the specimen and station and the set of equations can be solved. Usually additional observations are made and a least-squares solution obtained. The surplus equations afford an estimate of the experimental error in the observations. This makes it possible to test whether or not the observed differences between the stations exceed experimental errors and to attach an appropriate error to the values calculated for the specimens.

If a special symmetry is used in the assignment of specimens to stations, then improved precision and ease in solving the equations results. There are certain advantageous numbers of stations to place on a wheel because of the combinatorial properties of numbers. A simple case of a wheel with seven stations and seven specimens, A through G, will illustrate the principle involved:

	Station Number						
	1	2	3	4	5	6	7
Run No. a	A	B	C	D	E	F	G
Run No. b	B	C	D	E	F	G	A
Run No. c	D	E	F	G	A	B	C

Thus station 1 is occupied in turn by specimens A, B, and D and specimen A occupies stations 1, 7, and 5 in turn. Inspection shows certain relations have been achieved. The three stations that are occupied by A also encounter the six other members of the complete set B, D, C, G, E, F, of the other specimens. Thus A can always be compared with any other specimen occupying the same station.

Similarly station 1 which encounters the specimens A, B, and D can, by means of these specimens, be directly compared with all the other six stations. Specimen A permits station 1 to be compared with 5 and 7; specimen B compared station 1 with 2 and 6; and specimen D compared station 1 with 3 and 4.

Suppose we wish to evaluate station 1 in terms of the average performance of all seven stations. Let A_{1a}, B_{1b}, etc., represent the observation made on the specimen, A, B, C, etc., in the designated stations and runs. Consider the three observations on specimen A. These observations permit the comparison of station 1 with the average of stations 5 and 7. It is more convenient to multiply by 2 and write:

$$2A_{1a} - A_{7b} - A_{5c} = \Delta_{1,7,5}.$$

Similarly $$2B_{1b} - B_{2a} - B_{6c} = \Delta_{1,2,6}.$$

and $$2D_{1c} - D_{3b} - D_{4a} = \Delta_{1,3,4}.$$

Each equation is free of any specimen contribution. What about run effects? The run effects, if present, are designated by the letters a, b, and c. Observe that the sum of these three equations involves the subscripts a, b, and c each twice with a negative sign and twice with a positive sign. That is the run effects, if any, neatly cancel out, provided that conditions in each run are constant. We may, therefore, drop the a, b, and c subscripts and treat the differences as differences *between stations*, i.e.,

$$6[1] - [2] - [3] - [4] - [5] - [6] - [7] = \Sigma\Delta$$

where the station numbers are given in the brackets. We may add to this equation the equation

$$[1] - [1] = 0$$

which simply says that station 1 is equal to station 1 (with no error of measurement).

$$7[1] - \{[1] + [2] + \ldots + [7]\} = \Sigma\Delta.$$

Dividing by 7

$$[1] - \text{mean of all stations} = \Sigma\Delta/7$$

or

$$[1] = \text{mean of all stations} + \Sigma\Delta/7.$$

Customarily the "mean of all stations" is a number which is the average of all 21 observations. This gives equal weight to every station, every specimen and each run. The Δ's are obtained directly from the observations so that it is a simple matter to calculate a value for each station. These values are completely comparable because the specimen and run effects have been neatly removed making use of the special properties associated with the above triads of letters.

An exactly parallel procedure leads to estimates for each of the seven specimens, estimates that are corrected for any station differences. The simple sum of the seven observations for each run contains the contributions of all specimens and all stations so these sums may be compared directly to detect differences between runs.

If this procedure shows the stations to be satisfactorily equivalent there will be no need to follow any particular schedule in assigning specimens to stations and no need to make any adjustments. If there are important differences among the stations there is a choice of getting a better wheel or following a suitable scheme of specimen placement that will permit adjustment for station differences.

Clearly, if there are as many specimens as stations, making two runs leads to a unique solution for the differences, but without providing an estimate of the experimental error. In most instances it will be desired to hold the number of runs to three or four because the specimens have to be moved to new stations after each run. Several possible schemes using 3 or 4 runs are listed in table 1. An extensive collection of designs is available in a Bulletin [Bose, Clatworthy, and Shrikhande, 1954].

The example with seven stations just discussed is particularly simple in that any given specimen is

TABLE 1. *Examples of designs useful for intercomparing positions in apparatus*

8 Stations	9 Stations
R5, p. 185	SR12, p. 143
A B C D E F G H	A B C D E F G H I
B C D E F G H A	E C D B I H F G A
D E F G H A B C	F A E I G D B C H
10 Stations	**13 Stations**
T6, p. 231	C1, p. 250
A B C D E F G H I J	A B C D E F G H I J K L M
B H J A F C E I D G	C D E F G H I J K L M A B
E C D G H A I J B F	I J K L M A B C D E F G H
15 Stations	**16 Stations**
T28, p. 237	LS 14, p. 245
A B C D E F G H I J K L M N O	A B C D E F G H I J K L M N O P
J O K G F N E L D H A M I B C	O P M N K L I J F G H C D A B C
O G I L M D K G J E N A B H F	L I J K B C D A O P M N F G H E
19 Stations	
S1.1, p. 218	
A B C D E F G H I J K L M N O P Q R S	
C N I A K L F J Q S B M G P H E R O D	
B Q L E R N A I G F H K P O D J S C M	
12 Stations	**14 Stations**
R15, p. 188	R24, p. 192
A B C D E F G H I J K L	A B C D E F G H I J K L M N
B C D E F G H I J K L A	L M N H I J K E F G A B C D
E F G H I J K L A B C D	J K L M N H I C D E F G A B
G H I J K L A B C D E F	I J K L M N H B C D E F G A

Page numbers and design identification refer to: Bose, R. C., Clatworthy, W. H., and Shrikhande, S. S., Tables of Partially Balanced Designs with Two Associate Classes. North Carolina Agricultural Experiment Station Technical Bulletin No. 107 (1954).

paired just once with all the other specimens. By "paired" is meant "meets on the same station." This it not true for all the other designs listed in table 1. The arithmetical procedure for computing the estimates for specimens and stations for these designs is given in the above mentioned Bulletin. Above each design in table 1 is given the identification number and page reference where the design is listed in the Bulletin.

Certain of the designs show a simple cyclic displacement of the specimens for the successive runs. The order of the columns (stations) in the designs may be randomized and the rows run in any order without changing the properties of the design.

The apparatus described in this paper uses a wheel with 20 stations. We might use the design for 19 stations and leave one station on the wheel unfilled. An alternative was chosen by using a design for 10 stations and using this design twice. In effect this means two separate and independent sets of data and it was necessary to achieve some way to tie together all 20 stations, which was accomplished by interlacing the stations. First a pair from one design, then a pair from the second design and so on. This spread the two designs evenly over the whole wheel. The assumption was made that the 10 stations assigned to one design would have very closely the same average as the 10 stations assigned to the other design. When each station is rated as a ratio to the average for the set to which it belongs, the 20 ratios would fairly reflect the differences among all the stations.

A wheel with 25 stations could be filled with designs for 10 and 15 stations. By combining designs a wheel of any given number of stations may be accommodated.

The general availability of computers will probably mean that the matrix of equations will be solved with their help. The particular merit of these designs is that the solution can be obtained by inspection; thus consider the design for 10 stations given below:

Station Number									
1	2	3	4	5	6	7	8	9	10
A	B	C	D	E	F	G	H	I	J
B	H	J	A	F	C	E	I	D	G
E	C	D	G	H	A	I	J	B	F

The assignment of the specimens to stations makes it possible to intercompare the specimens without introducing the differences between stations should these be present. Consider specimen A which appears in stations 1, 4, and 6 along with specimens B, E; D, G; F, C. Direct comparisons of A with these six specimens (two at a time) is therefore possible staying within a station. Three other specimens, H, I, and J never share a station with specimen A. The object is to effect comparisons of A with H, I, and J without introducing station differences. We observe that stations 2 and 5 permit the comparison of H with B, C; E and F. Similarly stations 7 and 9 are used to compare I with B, D; E and G. Finally stations 3 and 10 provide the comparison for J with C, D; F and G. We may combine these three sets of comparisons and obtain the result that H, I, and J as a group may be contrasted with B, C, D, E, F, and G as a group.

It was shown above that stations 1, 4, and 6 provided the station-free comparison of A with B, C, D, E, F, and G as a group. We also have just obtained the station-free contrast of B, C, D, E, F, and G as a group with the group H, I, and J.

Therefore A can be compared with H, I, and J using the group, B, C, D, E, F, G, as an intermediary. Evidently A may be compared with all other specimens using only comparisons made within stations.

We have, therefore, the following comparisons:

$$2A - B - E$$

$$2A - D - G$$

$$2A - C - F$$

and

$$B + C - 2H$$

$$E + F - 2H$$

$$G + E - 2I$$

$$B + D - 2I$$

$$C + D - 2J$$

$$F + G - 2J$$

Note that by multiplying the first three comparisons by 6 and then summing them with the last six comparisons, we have as a result

$$36A - 4(B + C + D + E + F + G + H + I + J).$$

Adding and subtracting 4A gives

$$40A - 4 \text{ (total of all sources)}.$$

Dividing by 40 gives A – (average of all 10 sources) in terms of the differences. These operations are shown, for both sources and stations using actual counts, in tables 3 and 4.

Imagine for a moment a perfect wheel, all stations identical, also identical specimens, and identical runs. The 30 observations would then be identical except for experimental error. In an actual experiment each observation may be regarded as undergoing three displacements. The specimen, the station, and the run all combine to effect a net displacement.

The preceding paragraph indicates how to obtain the displacement contributed by specimen A. Using these predicted quantities, i.e., the least square estimates, a matching set of predicted expected values can be obtained for comparison with the actual observations. In fact, the sum of the squares of the 30 discrepancies between observed and predicted values is a measure of the experimental error.

The sum of the squares of the deviations must be divided by (30–1–9–9–2) or 9 to obtain the mean square error. The deductions from 30 refer to the mean, the nine independent specimen constants, the nine independent station constants and two independent run constants. The standard deviation of a single observation is obtained by taking the square root of the mean square error.

In the present experiment a wheel with 20 stations was being used to intercompare sources used as radioactivity standards. There is no suitable standard design for 20 stations with a limited number of interchanges for the sources. Consequently the design for 10 items with three interchanges was used twice. The 20 stations were interlaced by assigning stations 1, 2, 5, 6, 9, 10, 13, 14, 17, 18 to one design and the remaining 10 stations to the other design. This assumes that the averages for the two sets of 10 stations will each be representative of the wheel as a whole. This assumption can be verified when the data become available. All 20 stations can be put on a comparable footing by expressing each station as a percent of the average for the group of 10 to which it belongs; this assumes that the averages of the two groups of 10 stations are the same.

TABLE 2. *Counts minus one million for each of the three stations occupied by each source*

Station number	Run I		Run II		Run III	
	Source	Count *	Source	Count *	Source	Count *
1	K	42558	B	35323	O	42911
2	L	50654	R	40375	E	42384
5	O	42711	A	37296	L	49092
6	P	37720	K	44580	F	39822
9	A	40622	Q	35730	K	43096
10	B	36471	F	40506	R	41525
13	E	41432	P	40623	A	39876
14	F	39051	E	42361	Q	36443
17	Q	36856	L	49803	B	38637
18	R	41535	O	46438	P	41311
Total exp't 1		4096610		4130335		4155097
3	M	38417	D	37985	T	36523
4	N	35271	M	38203	J	35817
7	T	36910	N	38137	C	38110
8	J	43440	T	39107	G	37974
11	C	38773	J	37635	D	40225
12	D	37316	C	42288	H	37859
15	G	35733	J	36176	N	38121
16	H	37663	G	37996	M	35500
19	I	35491	C	38815	I	42263
20	J	37916	H	40813	J	34940
Total exp't 2		3769930		3871555		3773332

* Actual counts diminished by one million

Twenty sources, identified by letters, were assigned to the 20 stations as shown in table 2. Once the sources were assigned to the stations for the first run, the wheel was started and 5 min counts made at each station giving a count somewhat over 200,000. Five revolutions of the wheel were made without disturbing the sources. The five revolutions with short stops makes for a more equitable sampling of the background and machine performance during the time required for a run.

At the conclusion of the first run, the sources were transferred to new assigned stations and another five revolutions made. The sources were again shifted for the third run. The station assignments are such as to make possible the intercomparison of any station with the other nine stations in its group without introducing differences between the sources. Counts were recorded for each 5 min period. The five counts were summed and diminished by one million and the remainders entered in table 2. These coded values are all that is needed because the calculations involve differences between the entries in table 2. Naturally the raw data reflect the combined effects of sources and positions. Thus the simple average of the three A counts involves any effects associated with station 9, 5, and 13. Similarly the average of the three counts recorded for station 6 depends on the values for sources P, K, and F. The merit of the design rests in the ease with which the effects associated with individual station and sources can be disentangled.

Tables 3 and 4 show specimen computations for source A and station No. 6 in the first group of 10. The adjustment for a source is made up of quantities obtained by taking *differences* between sources *within* the same station. Station effects are therefore not present. Similarly, stations are evaluated by taking *differences* between stations using the *same* source, and source effects are thereby eliminated. As a datum, or reference point, the average of all 30 counts is used. The computed adjustments are added or subtracted from this grand average. This gives, on the one hand, adjusted estimates for sources as though there were no differences between wheel stations: and equally adjusted values for stations as though 10 identical sources had been available to compare the stations.

TABLE 3. *Calculation of adjustments to observed values for source, using source A as an example*

Station	Sources	
9	2A − Q − K = 81244 − 35730 − 43096 =	2418
5	2A − L − O = 74592 − 49092 − 42711 =	−17211
13	2A − E − P = 79752 − 41432 − 40623 =	−2303
	Total =	−17096
	Multiply total by six * =	−102576
1	K + O − 2B = 42558 + 42911 − 70646 =	14823
17	Q + L − 2R = 36856 + 49803 − 77274 =	9385
14	E + Q − 2F = 42361 + 36443 − 78102 =	702
6	P + K − 2L = 37720 + 44580 − 79644 =	2656
18	O + P − 2R = 46438 + 41311 − 83070 =	4679
2	L + E − 2R = 50654 + 42384 − 80750 =	12288
	Total below double line =	58043
	Divide by 40 = Adjustment =	1451
	Add grand average of 30 counts =	41258
	Adjusted value for A	99807

* The factor "six" is obtained by inspection to insure that each letter occurs equally often with a minus sign when the summation is made

TABLE 4. *Calculation of adjustments to observed values for stations using station No. 6 as an example*

Source	Stations
P	2[6] − [13] − [18] = 75440 − 40623 − 41311 = −6494
K	2[6] − [1] − [9] = 89160 − 42558 − 43096 = 3506
F	2[6] − [14] − [10] = 79644 − 39051 − 40506 = 87
	Total = −2901
	Multiply total by six = −17406
R	[18] + [10] − 2[2] = 41535 + 41525 − 80750 = 2310
E	[13] + [14] − 2[2] = 41432 + 42361 − 84768 = −975
O	[18] + [1] − 2[5] = 46438 + 42911 − 85422 = 3927
A	[9] + [13] − 2[5] = 40622 + 39876 − 74592 = 5906
Q	[9] + [14] − 2[17] = 35730 + 36443 − 73712 = −1539
B	[10] + [1] − 2[17] = 36471 + 35323 − 77274 = −5480
	Total below double line = −13257
	Divide by 40 = Adjustment = −331
	Add grand average of 30 counts = 41258
	Adjusted value for station No. 6 = 40927

TABLE 5. *Adjustments to station and source values and comparison with unadjusted values*

Station number	Station adjustment	Adjusted value	Unadj.[a] value	Source number	Source adjustment	Adjusted value	Unadj.[b] value
			Experiment I				
1	−1608	39650	40264	A	−1451	39807[d]	39265
2	+461	41719	44471	B	−4341	36917	36810
5	−1774	39484	43033	E	+499	41756	42059
6	−331	40927[c]	40707	F	−1570	39688	39793
9	−208	41050	39816	K	+2869	44127	43411
10	+555	41813	39501	L	+8785	50043	49850
13	+355	41613	40644	O	+3314	44572	44020
14	+91	41349	39285	P	−1957	39301	39885
17	+733	41991	41765	Q	−5120	36138	36343
18	+1727	42985	43095	R	−1027	40231	41145
Averages	.....	41258	41258	.....	.....	41258	41258
			Experiment II				
3	−371	37676	37642	C	+33	38080	38566
4	−1104	36943	36430	D	+557	38604	38509
7	+142	38189	37719	G	−860	37187	37234
8	+1228	39275	40174	H	+1161	39208	38778
11	+1024	39071	38878	I	−2906	35141	35536
12	−940	37107	39154	J	−1169	36878	37123
15	+77	38124	36677	M	+205	38252	37373
16	−1163	36884	37053	N	−576	37471	37176
19	+293	38340	38856	T	−867	37180	37513
20	+814	38861	37890	U	+442	42470	42664
Averages	.....	38047	38047	.....	.....	38047	38047

[a] The unadjusted value is the average of the three observed counts (table 2) for the station.
[b] The unadjusted value is the average of the three observed counts (table 2) on the source.
[c] Taken from table 4.
[d] Taken from table 3.

No adjustments are required for the run totals because the effects of all 10 sources and all 10 stations are present in every run. Unavoidably every one of the 30 counts is subject to the counting error and any unequalized drifts in background or counting electronics. The adjusted values shown in table 5 are the best estimates of source and station characteristics. We can use these adjusted values, together with the run averages, to compute an ideal table. In table 6 every actual count is replaced by an "ideal" value.

TABLE 6. *"Ideal" values calculated using best estimates for stations and sources*

Station No.	Run number I	Run number II	Run number III	Station No.	Run number I	Run number II	Run number III
1	42222	35355	43216	3	37527	38902	36495
2	50207	40738	42469	4	36013	37817	35460
5	42501	38079	48521	7	36968	38282	37908
6	38673	43842	39609	8	43344	39077	38101
9	39302	35976	44171	11	38750	38571	39314
10	37175	40289	41038	12	37310	42199	37954
13	41814	39702	40414	15	36910	35887	37234
14	39482	41893	36481	16	37691	36693	36775
17	36574	50822	37902	19	35080	39042	42449
18	41661	46345	41280	20	37338	40691	35641

The "ideal" values are obtained by combining the calculated adjustment for the station, the source and the run and adding the result to the grand average. The "ideal" value for the count obtained for source K in station 1 in run 1 is obtained by taking from table 5 the station adjustment (−1608); the source adjustment (+2869); the run adjustment (−297). The run adjustment is the difference between the grand average (table 5) and the run 1 average (table 2). The net adjustment, (2869−1608−297) or 964 when added to the grand average, 41258, gives the "ideal" value of 42222 for this observation. The discrepancies between the actual counts and these "ideal" values computed from the best estimates are a measure of the errors involved.

Table 7 lists the differences between the observed counts and the "ideal" values computed using the best estimates for sources, stations and runs. These best estimates impose 21 constraints on the data leaving nine degrees of freedom available for the estimation of error. The two error variances should be compared with the error variances listed in table 8 which were obtained by the computer using unrounded numerical values. The average count is about 1 040 000. Assuming the Poisson distribution the error variance should equal the mean count. Both estimates of error slightly exceed theory but are well within the limits that can be expected for estimates based on just nine degrees of freedom. Evidently the plan of work and equipment gave data which were close to the theoretical Poisson error.

The mean squares shown in table 8 provide the means for judging whether the data provide convincing evidence of differences among the wheel stations. The ratio of the mean square for adjusted positions to the error mean square is the familiar statistic F. This ratio is 2.24 for experiment 1 and 1.44 for experiment 2. Both ratios are less than the 90 percent value (2.44) tables for nine degrees of freedom for both numerator and denominator. The fact that both mean squares do exceed the error mean square does suggest there may be small differences among the stations too small to be conclusively detected in these experiments. If these possible station differences are ignored, there would result some small increase in the error variance associated with the source averages.

TABLE 7. *Differences between observed counts and calculated values shown in table 6*

Station No.	Experiment 1			Station No.	Experiment 2		
	Run number				Run number		
	I	II	III		I	II	III
1	−336	32	305	3	−890	917	−28
2	−447	363	85	4	742	−386	−357
5	−210	783	−571	7	58	145	−202
6	953	−738	−213	8	−96	−30	127
9	−1320	246	1075	11	−23	936	−911
10	704	−217	−487	12	−6	−89	95
13	382	−921	538	15	1177	−289	−887
14	431	−468	38	16	28	−1303	1275
17	−282	1019	−735	19	−411	227	186
18	126	−93	−31	20	−578	−122	701
Sum of squared differences		10 139 049			10 951 000		
Divide by 9 Error variance		1 126 561			1 216 778		

TABLE 8. *Mean squares from analysis of variance*

Variance source	Degrees of freedom	Mean square	
		Experiment I	Experiment 2
Runs	2	768 160	3 353 392
Unadj. stations	9	9 277 905	4 290 006
Adju.t. stations	9	2 526 719	1 754 738
Unadj. sources	9	46 227 069	10 549 753
Adjust. sources	9	39 475 883	8 014 484
Error variance	9	1 126 560	1 216 778

TABLE 9. *Analysis of variance ignoring stations*

Item	Degrees freedom	Mean square	
		Experiment I	Experiment II
Runs	2	768 160	3 352 392
Sources	9	46 227 069	10 549 753
Error	18	1 826 640	1 485 758

In fact if it be assumed that the sources were assigned at random to the stations, the analysis of variance would appear as shown in table 9. The small increase in the error variance results from not correcting for the very small differences between stations.

Another way to make clear the minor contribution to error made by stations is to look at the amount by which the adjusted count for a station differs from the average count for all stations. The "adjusted" counts are adjusted to allow for the fact that *different* sources were usually in different stations. The differences are shown as percentages in table 10 and plotted in figure 6. The differences are of the order of one tenth of a percent which is quite reasonable for the counts available. The graph gives just a hint of a region of high values and a region of low values.

TABLE 10. *Percent by which stations differ from average station*

Station and percent				Station and percent			
1	0.154	10	−0.053	3	0.035	12	0.091
2	−.044	13	−.044	4	.106	15	−.007
5	.170	14	−.009	7	−.014	16	.112
6	.031	17	−.070	8	−.118	19	−.028
9	.020	18	−.166	11	−.099	20	−.078

Further study of the mean squares in table 8 reveals a much larger mean square for sources in experiment I than in experiment II. Source *L*, which is 0.844 percent above the average of all sources is largely responsible. No other source differs as much as half a percent from the average source. The three largest deviations in experiment I are 0.844, 0.492, and 0.417. In experiment II the three largest deviations are 0.426, 0.279, and 0.113. Apparently experiment I happened to get the sources that deviated most from the average, whereas experiment II got sources that, on the whole, gave somewhat lower counts than those forming experiment I. This state of affairs is plainly revealed in figure 7. This is not to imply great variation among the sources. All but one of the 20 sources fell in the range of 1 035 000 to 1 045 000 for their counts. The unadjusted counts are very similar to the adjusted counts because there was so little difference among the stations. In no case is the difference between observed and adjusted count as much as 1000.

There remains a remark about the mean squares found for runs. If the total exposure time remained the same for each run and the counting apparatus maintained performance, then the mean square for runs should approximate the mean square for error.

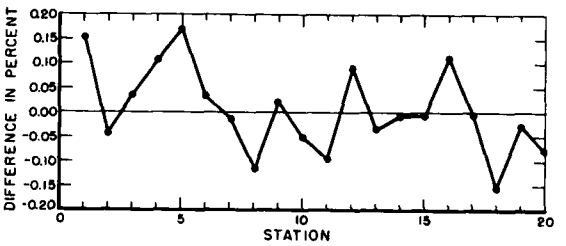


FIGURE 6. *Difference of each wheel station from wheel average, expressed in percent.*

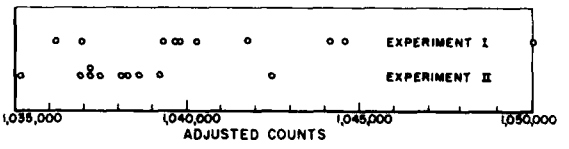


FIGURE 7. *Adjusted counts for sources.*

The somewhat larger mean square for runs in experiment II is without significance. The mean square would have to exceed the error mean square by a factor larger than four to suggest a real difference between runs.

The use of these "incomplete block" designs is not without a certain price. The original application of these designs was in agricultural field trials. If a large number of varieties of wheat are under comparison it is clear that a block of 20 plots requires a large area of ground. Some of the plots will be at considerable distances from each other and may encounter substantial differences in the soil. Experience showed that comparisons between widely separated plots are subject to greater errors than comparisons between nearby plots. The basic idea back of the incomplete block scheme was to take advantage of the very substantial reduction in experimental error that came from using small blocks. The reduction in error far outweighed the additional mathematics. The indirect comparisons are not as effective as direct comparisons, and therefore result in a lower efficiency. The efficiency of the design used in this work is approximately 70 percent. This may be translated into the following terms. The standard error for the average of three counts with the block design is about that which would be associated with the average of two counts without this design.

In agriculture the sizable reductions in error which resulted from using small blocks outweighed the loss in efficiency. The present experiment affords an interesting example where the reduction in error achieved by eliminating position contributions is relatively slight. On the basis of the error variances given in tables 8 and 9 the variance is increased from 1.17 to 1.66 million when the position effect is left in. Dividing 1.17 by two and 1.66 by three gives 0.586×10^6 and 0.552×10^6, respectively, as the variance for the source averages. All this effort would appear to have been to no avail.

One important consequence did come from the use of the design. The design made it possible to evaluate the station effects using the *same* data that were collected to calibrate the sources. Evidence was obtained that the wheel stations are very closely identical. Actually there is no need to take account of wheel stations unless considerably greater counts are taken. In that event the contribution arising from station differences will be relatively more important. It should be pointed out that if the stations had differed by about as much as the source, the precision gained by correcting for station effects would have been impressive. Obviously if stations differed as much as sources, discrimination between the sources becomes impossible. In this event the adjustment for source effects would save the day *provided* a design was used that makes such an adjustment possible.

An exacting test was made of the effectiveness of the numerical adjustments by purposely introducing substantial biases into the wheel stations. Single cardboard shims were placed under the sources (fig. 4) on five of the 20 stations, so as to increase the source-to-detector distance. Two shims were placed on five other stations, three shims on still another five stations and the remaining five stations were left without shims. The stations were picked at random in allocating the shims. The shims stayed on the stations throughout the experiment.

Twenty sources were placed on the wheel and the same procedure used as before. In this case three revolutions of the wheel constituted a run. The average count per source (and station) *per run* (sources remaining in their stations) was 318391. The average total count per source (and station) for three runs was three times 318391, or 955173.

TABLE 11. *Comparison of sources using biased wheel*
Each source and station expressed as a ratio to the average source and station.

Section A					Section B			
	2	3	4	5	6	7	8	9
Source	No bias	Biased stations		Diff. *Percent*	Station No.	Biased stations		Diff. *Percent*
		Exp't I	Exp't II			Exp't I	Exp't II	
K	1.0006	1.0000	1.0024	−0.24	1	1.0198	1.0219	−0.21
L	1.0016	0.9998	1.0030	−.32	2	0.9827	0.9810	.17
O	1.0005	1.0002	0.9964	.38	5	1.0081	1.0087	−.06
P	0.9997	1.0012	.9986	.26	6	1.0087	1.0055	.32
A	1.0029	1.0073	1.0065	.08	9	0.9795	0.9818	−.23
B	0.9989	1.0000	0.9987	.13	10	1.0064	1.0084	−.20
E	1.0015	0.9996	1.0020	−.24	13	0.9928	0.9905	.23
F	0.9980	.9972	0.9992	−.20	14	1.0064	1.0070	−.06
Q	1.0005	.9984	.9981	.03	17	1.0210	1.0190	.20
R	1.0014	1.0014	1.0011	.03	18	0.9797	0.9821	−.24
M	0.9987	0.9990	1.0005	−0.15	3	1.0182	1.0173	0.09
N	.9986	.9995	0.9987	.08	4	0.9923	0.9927	−.04
T	1.0029	1.0059	1.0043	.16	7	.9806	.9812	−.06
U	0.9966	0.9970	0.9966	.04	8	1.0184	1.0165	.19
C	.9964	.9977	.9983	−.06	11	0.9952	0.9942	.10
D	1.0011	1.0016	.9978	.38	12	.9919	.9924	−.05
G	0.9992	1.000	.9993	.14	15	1.0173	1.0154	.19
H	.9968	0.9956	.9965	−.09	16	1.0055	1.0076	−.21
I	1.0041	1.0020	1.0035	−.15	19	0.9815	0.9848	−.33
J	1.0000	0.9957	0.9985	−.28	20	.9940	.9920	.20

The above experiment was repeated and the relative values of sources and stations computed. Table 11 lists the results of these computations. The entries in section A of the table show each source as a ratio to the average source and in section B show each station as a ratio to the average station. The difference between the stations with no shims and those with three shims is nearly 4 percent. In spite of these biases introduced into the wheel the adjusted values of the sources (col. 3 and 4) agree with the ratios obtained in another trial using the wheel without shims (col. 2). No adjustments were made for the ratios in column 2, the wheel stations being assumed to be without bias. In fact very slight biases do exist as shown in the preceding study.

The average magnitude of the twenty *differences* between the paired estimates for the sources is 0.172 percent and for the stations is 0.169 percent. Each estimate is based on about 950 000 counts. As stated earlier, the price of using the experimental design that makes possible the adjustment for the effect of stations, is a certain loss in efficiency. In this case the efficiency is about 70 percent so that the effective

count is 950 000 × 0.70 or 665 000. The square root of 665 000 is 816, therefore the *expected* standard deviation of an estimate of a source is 816/665 000 or 0.123 percent. The expected *average difference* between two measurements each with standard deviation 0.123 is obtained by multiplying by $2/\sqrt{\pi}$ or 1.128. The theoretical average difference, $0.123 \times 1.128 = 0.14$ is only slightly less than the experimental average difference.

The good concordance between experiments I and II confirms the error as *calculated* from the statistical analysis on the separate experiments. These errors were 0.15 and 0.14 percent, respectively. The evaluation of the sources is confirmed by the two experiments and the evaluation of the experimental error is also confirmed by the paired comparisons.

Because sources are compared by taking ratios of counts, the whole statistical analysis was repeated using the logarithms of the observed counts. The analysis of variance and the adjustments in the first analysis were made using differences rather than ratios, because of the near identities of both sources and stations. The analysis using logarithms *did not* alter any of the conclusions. Fortunately the counts were large and varied over a very small range. Over this range the logarithms are acceptably proportional to the counts so that the effect of using logarithms was just that of changing units.

6. References

Bose, R. C., Clatworthy, W. H., and Shrikhande, S. S. (1954), Tables of partially balanced designs with two associate classes, North Carolina Agricultural Experiment Station Technical Bulletin No. 107.

DeWaard, H. (1955), Stabilizing scintillation spectrometers with counting-rate-difference feedback, Nucleonics **13**, July, p. 36.

Hutchinson, J. M. R. H. (1960), Calibration of five gamma-emitting nuclides for emission rate, NBS Tech. Note 71.

Mann, W. B., and Seliger, H. H. (1958), Preparation, maintenance and application of standards of radioactivity, NBS Circ. 594.

(Paper 70C2-219)

Reprinted from SCIENCE, October 22, 1954, Vol. 120, No. 3121, pages 627–631.

Instrumental Drift*

W. J. Youden

National Bureau of Standards, Washington 25, D.C.

THE developments in instrumentation and control devices in recent years are manifest in most laboratories. These advances have brought better measurements and have eased the labor of obtaining and recording them. Furthermore, improved instrumentation often has made it fea ible to take more measurements. There is another consequence, one that many experimenters will consider an advantage, to be credited to better instruments. Better measurements, and more of them, have made it possible to interpret most data without recourse to statistical techniques.

Experimenters habitually try to select instruments and to control measurement procedures in order to get reproducible measurements that are good enough for their immediate purposes. These purposes generally fall into two classes: either the experimenter wants to keep the uncertainty in the result below some specified value, or else he wants to be able to distinguish between objects if these differ by some minimum amount in the measured property. If the worker succeeds in these respects, the interpretation of the data is simplified, because the uncertainties in the measurements can be, and usually are, ignored.

Apparently it is easier for many people to obtain elaborate and expensive control devices than it is to delve into the subject of the statistical design of experiments. Or they may be unaware that statistical design can bring the same kind of improvement in the data that comes from providing a uniform environment and will do this with little or no expense. The ideal measurement procedure should give results that the experimenter can accept without worrying about their reliability. The experimenter is then free for the task of studying the relationships tnat are involved in his scientific problem. In most cases the measurements are subject to random and other unknown sources of error that may either obscure relationships or even give the appearance of relationships when in fact there are none.

A good place to introduce statistics is in the preliminary trials an experimenter makes to assure himself that his apparatus and instruments are in a satisfactory operating condition. Consider the question of whether or not the instrument is subject to drift. Drift is usually explored by making a series of repeated measurements on the same object. Another question then plagues the worker. How can these repeated measurements be made independent of one another? How can the operator "forget" previous readings so that subsequent readings will not be influenced by earlier ones? These matters will be considered later.

Suppose the experimenter has made a series of measurements on the same object and has plotted the values as ordinates against the serial numbers of the measurements. A line drawn parallel to the x-axis with y equal to the average of all the readings will provide a visual test to detect trends in the sequence of readings. The experimenter would like to have the measurements indiscriminately scattered about the line and confined between two bracketing parallel lines as close as possible to the average line. If there is a pronounced trend, the visual test will reveal it. On the other hand, the experimenter may not be sure. Here, then, is the opportunity to use an objective statistical criterion to bolster his own judgment.

Table 1 lists measurements $y_1, y_2, \ldots, y_n$ in the order in which they were obtained. Two quantities, S^2 and D^2, may be computed from the observations in Table 1. The ratio of D^2 to S^2 should fall within predictable limits about the integer 2 if the results are free from trends. The quantity S^2 is the sum of the squares of the deviations of the plotted points from the horizontal line through the average. The formula

$$S^2 = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n},$$

where $\bar{y}$ is the average value, is a convenient way to obtain this sum of squares. Incidentally, the estimate of the standard deviation for these measurements is $\sqrt{S^2/(n-1)}$.

The quantity D^2 is the sum of the squares of the differences between successive measurements: $D^2 = \Sigma d_i^2$. It will be noted that the interval between two *successive* measurements gives only slight opportunity for the trend to operate. The d's are nearly what they would be if there were no trend at all. In contrast, the deviations between the individual y's and $\bar{y}$ are susceptible to the trend, and S^2 will be larger than it would otherwise be. The value of the ratio D^2/S^2 will then fall below 2.0.

It remains to set up some criterions for the allowable ratio of D^2/S^2. In any set of observations, the

Table 1. Successive differences between measurements.

Order of measurement	Measurement	Successive difference
1	y_1	$d_1 = y_1 - y_2$
2	y_2	$d_2 = y_2 - y_3$
3	y_3	
•		
•		
•		
$n-1$	y_{n-1}	$d_{n-1} = y_{n-1} - y_n$
n	y_n	

errors of measurement may, by chance, fall into suspicious configurations even when there is no trend. This is more likely to happen if the series is a short one so the limits for the ratio D^2/S^2 will depend on the number n of measurements.

Bennett (*1, 2*) has adapted some tables, published by Hart (*3*), that list limits of D^2/S^2, each of which will be exceeded on the average in 1 out of 20 sequences (or 1 in 100) for sequences that are not afflicted by any trend whatsoever. Thus, if a particular sequence does transgress these boundaries, it is usual to consider this as evidence of a trend rather than as a very improbable occurrence. Table 2 shows some specimen values of the limits taken from Bennett's table.

Sufficiently low values of the ratio D^2/S^2 are evidence of a trend. Overly large values of D^2/S^2 also indicate that the data depart from the expected random scatter. One way that the ratio may be inflated is by changing the zero setting or making other adjustments between successive readings. In general these adjustments will lead to a succession of large differences between successive readings and therefore will inflate D^2.

The following 20 determinations of the percentage of nickel were made on 20 successive segments of a rod of alloy by a spectrochemical procedure: 42.4, 40.8, 41.0, 41.8, 40.3, 40.8, 40.8, 39.6, 41.5, 41.5, 40.2, 40.4, 41.0, 42.2, 39.4, 41.0, 41.4, 40.6, 42.4, 40.8. It was important to know whether there was a trend along the rod. The computation for D^2/S^2 gave 31.32/12.99, or 2.41. The quotient is well within the listed limits for the ratio with n equal to 20, and there is no convincing evidence for a trend. The scatter of the data about the average line is shown in Fig. 1.

One obvious way to avoid the effect of remembering previous readings, referred to earlier, is to change the object being measured. At first thought, this would appear to make it impossible to detect any trend or drift in the measuring equipment. Certainly each reading will now depend on which object is measured and, if there is a drift, where the measurement is in the series. Such entanglement of effects can, however, be readily resolved if the objects are measured in an appropriate sequence. The devising of these sequences is one of the activities in the field of statistical design.

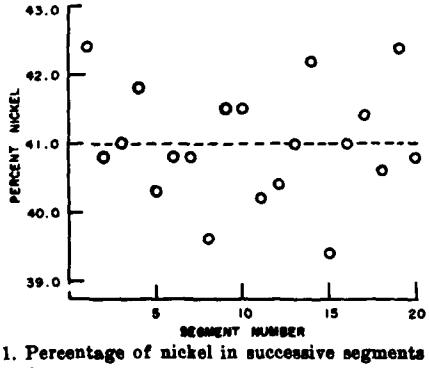


Fig. 1. Percentage of nickel in successive segments of an alloy rod.

Five objects, *A, B, C, D,* and *E,* may be available, and each could be measured four times in a sequence of 20 measurements. The problem is to set up a schedule that will still make it possible to detect the trend. Obviously nothing will be gained if four measurements are made on *A,* then four measurements on *B,* and so on. The memory difficulty is still present, and the values obtained for each object are inseparably combined with the drift, if any, in the instrument.

An alternative arrangement that begins to get into the problem is one that divides the sequence into four parts, each part containing all five objects. Thus,

BAEDC | BEDAC | EDCBA | CAEDB

The order of the objects within each part should be random. Now the average for the five objects in a particular block should be the same as the average in any other block except insofar as a trend happens to be present. In a coarse way, these averages, when plotted opposite 3, 8, 13, and 18, begin to reveal any instrument trend. The actual trend in any block that would be revealed by five ordinates is replaced by the average of these ordinates and centered in the middle of the block.

A modification of the afore-mentioned procedure will delineate the presumably rather smooth curve that corresponds to the true trend line during the measurements. The curve can be approximated by drawing short horizontal lines in a stepwise fashion along the curve. Each short horizontal line replaces the slant and slightly curved line in its vicinity. This horizontal line is located at a height equal to the average ordinate of the curve in the narrow band covered by the curved short line. If there were some way to determine the position of these short horizontal lines, the curve, or trend line, would stand revealed. It is better to have as many short lines as possible and to have them as short as possible. Ten short lines, each covering two measurements, afford a better approximation to the trend curve than four lines, each covering five measurements.

A difficulty then arises in the fact that the pair of objects used in any part will not be the same as the pair used in some other part of the curve. This would appear to make the averages for each pair useless for comparison, because the objects are different. If the pairs are formed in an appropriate manner, there is a simple procedure for comparing the parts, despite the fact that different objects occur in the different parts.

Five objects can be used to form 10 different pairs, each object appearing in four of the pairs.

Part	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
Object	*AB*	*DE*	*BC*	*EA*	*CD*	*EB*	*AC*	*BD*	*CE*	*DA*

These pairs break the trend curve into 10 parts. The order of the pairs is immaterial. The purpose is to determine the average values of the ordinates for each of the 10 parts, just as if all the measurements had been made on one object.

First use is made of the fact that the objects in any part, say *A* and *B* in part *a,* appear in six other parts. Thus, by using object *A,* the differences between part

Table 2. Limits for the ratio D^2/S^2.

No. in series n	1 in 20		1 in 100	
	Lower	Upper	Lower	Upper
5	0.82	3.18	0.54	3.46
10	1.06	2.94	0.75	3.25
15	1.21	2.79	0.92	3.08
20	1.30	2.70	1.04	2.96

Table 3. Determination of average value of ordinate a.

Using object	Difference between ordinates
A	$3(a-d) = x_1$
A	$3(a-g) = x_2$
A	$3(a-j) = x_3$
B	$3(a-c) = x_4$
B	$3(a-f) = x_5$
B	$3(a-h) = x_6$
C	$(c+g)-(e+i) = x_7$
D	$(h+j)-(b+e) = x_8$
E	$(d+f)-(b+i) = x_9$
Sum	$18a-2(b+c+d+\ldots+j) = \Sigma x_i$
Equivalently	$20a-2(a+b+c+\ldots+j) = \Sigma x_i$
And	a – (average ordinate over all parts) $= \Sigma x_i/20$

Table 4. Comparison of actual and calculated instrument drift.

Reading number	Instrument drift	Object and its value		Observed reading	Part	Calculated deviation from mean drift	Calculated drift*
1	0	A	75	75	a	– 2.9	2.1
2	4	B	85	89			
3	7	D	55	62	b	3.7	8.7
4	10	E	45	55			
5	13	B	85	98	c	8.5	13.5
6	15	C	65	80			
7	17	E	45	62	d	12.0	17.0
8	17	A	75	92			
9	16	C	65	81	e	10.7	15.7
10	15	D	55	70			
11	13	E	45	58	f	6.0	11.0
12	10	B	85	95			
13	7	A	75	82	g	0.5	5.5
14	4	C	65	69			
15	0	B	85	85	h	– 6.2	– 1.2
16	– 3	D	55	52			
17	– 7	C	65	58	i	– 14.1	– 9.1
18	– 10	E	45	35			
19	– 13	D	55	42	j	– 18.2	– 13.2
20	– 15	A	75	60			

* It is possible to determine this calculated drift only when the mean value of drift is known or determinable. In this example, the value 5.0 is used for the average of all ordinates from the curve of Fig. 2.

a and parts d, g, and j can be estimated; by using object B, the differences between part a and parts c, f, and h can be estimated. This leaves parts b, e, and i to be considered. Notice that, by using object C, parts c and g can be compared with parts e and i; by using object D, parts h and j can be compared with parts b and e; and finally E gives parts b and i in terms of parts d and f. The lower-case letters are used to represent the average ordinates of the parts. These differences are shown in Table 3.

When a result located in a part, say d, is subtracted from a result in another part, say a, using the *same* object (A), the value of A, whatever it may be, drops out. The first six differences tabulated are multiplied by 3 to bring the sum to the form shown in Table 3. All letters other than a have the coefficient – 2. The ordinate for part a, multiplied by 18, has twice the sum of the ordinates for all other parts subtracted from it, and Σx_i is the result. The difference is unchanged if twice ordinate a, or $2a$, is added and subtracted. Division by 20 then gives the ordinate for a when added to the *average* ordinate over all parts.

A constructed example illustrates how well the scheme works. Suppose an instrument drifts as shown in Fig. 2. The curve shows the drift expressed in units of the terminal figure recorded. The instrument starts out and drifts so that after a time the readings are too high by about 17 units in the last place; then the trend reverses and drops until at the end readings are too low by about 15 units.

Imagine that five objects, A, B, C, D, and E, are available and that these, when measured, should give the values 75, 85, 65, 55, and 45, respectively. By reading from the drift curve and by assigning the values for the objects, one obtains a sequence of 20 readings, as shown in the fourth column of Table 4.

The only information that is assumed available for the statistical analysis is the column of observed readings together with the identities of the objects. It is assumed that the objects themselves do not change in value during the observations. The calculation of the average drift corresponding to part a, using the data of Table 4 and the equations of Table 3, is shown in Table 5.

Table 5. Calculation of average drift corresponding to part a.

Using object	Difference between ordinates
A	$3(75-92) = -51$
A	$3(75-82) = -21$
A	$3(75-60) = 45$
B	$3(89-98) = -27$
B	$3(89-95) = -18$
B	$3(89-85) = 12$
C	$(80+69)-(81+58) = 10$
D	$(52+42)-(62+70) = -38$
E	$(62+58)-(55+35) = 30$
Sum	$18a-2(b+c+d+\ldots+j) = -58$
Equivalently	$20a-2(a+b+c+\ldots+j) = -58$
And	a – (average ordinate over all parts) $= -58/20$
Therefore calculated deviation from mean drift $= -2.9$	

Table 6. Determination of average value of ordinate b.

Using object	Differences between ordinates	
	In symbols	Using data of Table 4
D	$3(b-e)=x_1$	$3(62-70)=-24$
D	$3(b-h)=x_2$	$3(62-52)=30$
D	$3(b-j)=x_3$	$3(62-42)=60$
E	$3(b-d)=x_4$	$3(55-62)=-21$
E	$3(b-f)=x_5$	$3(55-58)=-9$
E	$3(b-i)=x_6$	$3(55-35)=60$
A	$(d+j)-(a+g)=x_7$	$(92+60)-(75+82)=-5$
B	$(f+h)-(a+c)=x_8$	$(95+85)-(89+98)=-7$
C	$(e+i)-(c+g)=x_9$	$(81+58)-(80+69)=-10$
Sum	$18b-2(a+c+d+\ldots+j)=\Sigma x_i$	$=74$
Equivalently	$20b-2(a+b+c+\ldots+j)=\Sigma x_i$	$=74$
And	b – (average ordinate over all parts) $=\Sigma x_i/20$	$=74/20$
		Calculated deviation from mean drift $=3.7$

To calculate the ordinate for b, we must set up another series of differences similar to the series used for the calculation of a (Table 3). These new differences are given in Table 6.

In setting up the series, note that the objects appearing in part b are objects D and E. Therefore the first 3 differences (x_1, x_2, x_3) are obtained by taking the value of object D in part b and subtracting from it the respective values of object D in the other three parts in which it appears. The next three differences (x_4, x_5, x_6) are obtained using the values of E in similar fashion. The difference x_7 is obtained by taking the sum of the values of A in the two parts where A appears with D and E and subtracting the sum of the two values of A that appear with B and C.

Similar sets of differences must be set up for all 10 parts in order to calculate the ordinates. In each instance the sum of all nine equations will be of the form shown in the sets given for a and b and, therefore, will provide a check that the proper differences have been set up. As a further check, when all 10 values of deviation from mean drift have been calculated, their sum should equal 0.

The numerical procedure outlined in the preceding paragraphs, leads to the estimates shown in the last column of Table 4. These, unavoidably, apply to both observations in the pair to which they are attached. Inspection reveals that the calculated drift is in excellent agreement with the averages of the two drifts recorded for each pair in the second column. Furthermore, by the pattern of deviations from mean drift (Fig. 2), the drift of the instrument stands revealed through the overlay of the different objects measured.

The drift curve was plotted on the assumption that the 20 observations were taken at equal intervals of time. This restriction may be relaxed, provided that the two observations forming any pair are taken in close succession and provided that the times are recorded. The x-axis becomes a time scale and the average ordinate for each part is located at the average time for the two observations.

One of the merits of using different objects is the fact that the observer cannot anticipate the next reading and this assists in the attainment of objectivity in the readings. This objectivity is particularly desirable in the matter of estimating the precision of the readings. Precision is usually estimated from immediately successive readings on the same object, and it is difficult to avoid forming an optimistic appraisal of the precision. The present scheme also makes possible an estimate of the precision. The numerical details are available (4–7).

So far all the emphasis has been placed on the performance of the instrument. The instrument will be used to measure objects, and it is reasonable to inquire whether the 20 observed readings in Table 4 can also be used to estimate the values of the five objects.

The pairs were formed in all possible ways from the five objects. Consequently, any given object has been matched with the four others in some four of the 10 parts. And, most important, in any part made up of two readings it can be assumed that the instrument drift error is approximately the same for each reading. In taking the difference between the readings for two objects in a part, the instrument drift, whatever in may be at that time, virtually drops out. The difference obtained is just about what it would be if there were no drift at all.

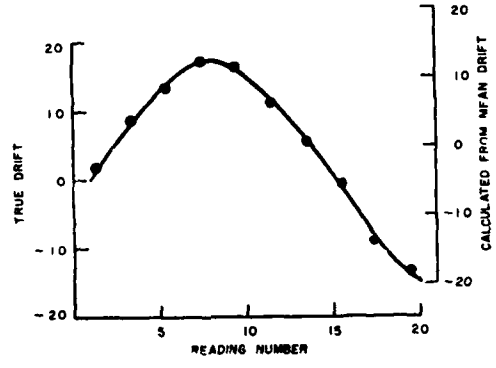


Fig. 2. Instrument drift in units of the terminal figure recorded.

Table 7. Comparison of correct and calculated values for the objects measured.

Object	Correct value	Calculated value	Calculated value less 5.0
A	75.0	79.4	74.4
B	85.0	90.4	85.4
C	65.0	70.6	65.6
D	55.0	59.0	54.0
E	45.0	50.6	45.6

The arithmetic for evaluating the objects is less involved than that used for the drift. To calculate an average for *A*, form the following differences:

Using part *a*, $A - B = -14$
Using part *g*, $A - C = 13$
Using part *j*, $A - D = 18$
Using part *d*, $A - E = 30$
Sum, $4A - (B + C + D + E) = 47$;
Equivalently, $5A - (A + B + C + D + E) = 47$;
And $A -$ average of all $= 9.4$;
Average of all 20 readings $= 70.0$; $A = 70.0 + 9.4 = 79.4$.

Table 7 shows the calculated averages for the objects alongside the correct values. There is evidently a marked discrepancy between the correct and calculated values. The fourth column shows the calculated values all diminished by 5.0, and now the two sets show good agreement. The correction, 5.0, cannot be evaluated in any actual case. It is, in fact, the average value of the drift introduced by the instrument. There is no way, short of the good fortune in having one of the objects a known standard, to separate out the *average drift from the average* of all the objects.

In much experimental work the *difference* between test items is all that is important to establish. Where absolute values are required, a standard object is indispensable. If the absolute value of one object is known, all other objects can then be determined.

Many choices are available in the construction of the sequence used. The parts or blocks may be of any size. For example, seven objects can be arranged in seven triads, or 10 objects in 10 triads.

ABD | BCE | CDF | DEG | EFA | FGB | GAC
ABE | HIJ | BHC | GEI | IDB | EFH | CJD | JGF | DAG | FCA

The first of these sequences is an example of a class of designs called balanced incomplete blocks. The second sequence is a partially balanced incomplete block design. Various discussions of these designs are available (*4, 5, 7*).

There is a final important *comment to make*. Comparisons of objects can be made even with a drifting instrument. Even when the instrument has been operating satisfactorily, the experimenter perforce usually has had to *assume* that this state was maintained while making the critical measurements. Statistical design makes it possible to show that the instrument did stay in adjustment and, if not, to introduce appropriate adjustments.

References and Notes

* Based on a talk given at the Gordon Research Conference on Instrumentation in 1954.
1. C. A. Bennett, *Ind. Eng. Chem.* **43**, 2063 (1951).
2. ——— and N. L. Franklin, *Statistical Analysis in Chemistry and the Chemical Industry* (Wiley, New York, 1954), p. 677.
3. B. I. Hart, *Ann. Math. Statistics* **13**, 445 (1942).
4. R. C. Bose and T. Shimamoto, *J. Am. Statistical Assoc.* **47**, 151 (1952).
5. W. G. Cochran and G. M. Cox, *Experimental Designs* (Wiley, New York, 1950).
6. W. S. Connor and W. J. Youden, *J. Research Natl. Bur. Standards* **53**, R. P. 2532 (1954).
7. O. Kempthorne, *The Design and Analysis of Experiments* (Wiley, New York, 1952).

Journal of Research of the National Bureau of Standards Vol. 53, No. 5, November 1954 Research Paper 2544

Comparison of Four National Radium Standards

Part 1. Experimental Procedures and Results

T. I. Davenport, W. B. Mann, C. C. McCraven, and C. C. Smith

Part 2. Statistical Procedures and Survey

W. S. Connor and W. J. Youden

Part 1

The two United States primary radium standards have been compared with the British primary radium standard and the Canadian national radium standard (1) by an ionization method, using the NBS standard electroscope, (2) calorimetrically, using the Peltier-cooling radiation balance, (3) by means of a Geiger-Müller counter, and (4) using a scintillation counter. Where there is little or no difference in gamma-ray source self-absorption, the four methods should, and in fact do, give good agreement. In the case of the Canadian national radium standard the difference in the results obtained is an indication of a difference in source self-absorption.

1. Introduction

During January and February 1954 the British primary radium standard and the Canadian national radium standard were at the National Bureau of Standards for the purpose of comparing these standards with the two United States primary radium standards at the Bureau. The intercomparisons were conducted over a period of 12 days and were made as exhaustive as possible, using the NBS electroscope, a Peltier radiation balance, and Geiger-Müller and scintillation counters.

2. Historical Background

In August 1911 Mme. Pierre Curie prepared, in Paris, a primary radium standard consisting of 21.99 mg of the pure anhydrous radium chloride that had been used to determine the atomic weight of radium as 226.0. This 21.99 mg of radium chloride was sealed into a glass tube 32 mm long, having an internal diameter of 1.45 mm and a wall thickness of 0.27 mm.

At the same time Professor Otto Hönigschmid, in Vienna, made three radium-standard preparations from very pure radium chloride consisting of 10.11, 31.17, and 40.43 mg of radium chloride sealed in glass tubes about 32 mm long, having internal diameters of 3.0 mm and wall thicknesses of 0.27 mm, each tube having a platinum wire sealed in one end. This wire was presumably to prevent the accumulation of static charge within the tubes. The purity of the radium chloride was defined by a radium atomic-weight determination, resulting in a value of 225.97. Of these the 31.17-mg preparation was chosen as a secondary standard. Mme. Curie's 21.99-mg primary standard and Professor Hönigschmid's secondary standard are generally and respectively referred to as the 1911 Paris and Vienna radium standards.

In 1934, after 23 years had elapsed, some concern was felt lest the Paris primary standard, together with a number of secondary radium standards, might explode on account of the accumulation of helium and chlorine and possible devitrification of the containing tubes. Hönigschmid was at that time carrying out, in Munich, a further determination of the atomic weight of radium, and accordingly the International Radium Standards Commission asked him to prepare new standards, using the same salt as for the atomic-weight determination.

For his atomic-weight determination, which was carried out in the early part of 1934, Hönigschmid used approximately 4 g of radium chloride, containing 3 g of radium element, that had been placed at his disposal by the Union Minière du Haut Katanga. This salt was purified by Hönigschmid to a point where spectroscopic analysis by Gerlach showed a maximum of 0.002 to 0.003 percent of barium atoms. A value was obtained for the atomic weight of radium equal to 226.05, which is currently accepted.

Hönigschmid then used some 817 mg of this highly purified anhydrous radium chloride to prepare 20 new standards of radium. Exactly who asked him to do this is not now quite clear. According to Mlle. Chamié [1],[1] the International Radium Standards Commission, at the suggestion of Stefan Meyer, "entrusted Mr. O. Hönigschmid with the preparation of 20 standards, using the salt he had purified and used in measuring the atomic weight of radium." According to Hönigschmid himself, however, in a paper [2] presented after his death by Stefan Meyer, the 20 standards were prepared "at the wish of the Belgian radium company." These two versions are, however, not irreconcilible if one assumes that the suggestion of the Belgian company was made known to the International Radium Standards Commission, which then gave it its official sanction.

[1] Figures in brackets indicate the literature references on page 272

The 20 new Hönigschmid standards were sealed into glass tubes on June 2, 1934, the glass tubing being similar to that used to seal the 1911 Vienna standard and having an internal diameter of 3.0 mm and a wall thickness of 0.27 mm. A platinum wire was sealed into the end of each standard.

One of the new Hönigschmid standards that was 42 mm long and contained 22.23 mg of radium chloride was selected as the new international standard, and its value was carefully compared with the 1911 Paris standard by gamma-ray measurements over a period of 4 years [1]. The Hönigschmid reference number for this standard is 5430. Hönigschmid states [2] that the error of a single weighing was not more than 0.02 mg. The gamma-ray comparison with the 1911 Paris standard showed a discrepancy, however, of 0.2 percent, corresponding to a weight of 22.27 mg as of June 2, 1934.

The first United States radium standard was brought to America in 1913 by Mme. Curie. This source contained 20.28 mg of radium chloride and was designated by the International Radium Standards Commission number IV (Vienna No. 6).

In 1936 two of the twenty Hönigschmid preparations were acquired as the United States primary radium standards. They are each designated by two numbers, namely, 5437, XIV and 5440, XV. The arabic numerals are those given by Hönigschmid, and the roman numerals are those assigned by the International Radium Standards Commission and imply that the standards have undergone gamma-ray comparison with the 1911 Paris and Vienna standards. The lengths of these two United States standards are 36 and 37 mm, and they contained 50.22 and 26.86 mg, respectively, of radium chloride as weighed by Hönigschmid on June 2, 1934. These weights correspond to 38.23 and 20.45 mg of radium element. The weights derived from a comparison with the Paris and Vienna 1911 standards corresponded, however, to only 38.13 and 20.38 mg, respectively, of radium element, as of June 1934.

The British primary radium standard is designated by one number only, namely, 5432. It is solely a standard by weight and was not compared with the 1911 Paris and Vienna standards. It is, however, one of the original Hönigschmid preparations sealed on June 2, 1934. Its length is 38.8 mm, and its salt content corresponds to 15.60 mg of radium element, as of that date. This standard replaced the first British radium standard, which had been in the custody of the National Physical Laboratory since 1913. This earlier standard was designated by the International Radium Standards Commission number III (Vienna No. 3).

The United States and British primary radium standards, as can be seen from figure 1, have low ratios of volume of salt to volume of tube. It is therefore to be expected that with the standards in a horizontal position and the grains of radium chloride distributed evenly along the tube their gamma-ray source self-absorption would be very nearly the same.

The Canadian national radium standard is however shorter and of smaller diameter than the Hönigschmid preparations, and it is tightly packed (fig. 1). It was sealed in June 1930 by the Union Minière du Haut Katanga, its contents, and that of six other sources in the custody of the National Research Council, having been taken from two tubes of radium chloride that had been prepared by the Union Minière in June 1924. Its weight was derived by gamma-ray comparison in 1933, in Paris and Vienna, with the 1911 standards, and it is designated by the number XIII. It is understood that no corrections for possible differences in self-absorption were made in these gamma-ray comparisons. Its length is 10.5 mm, its internal diameter 1.5 mm, and its salt content corresponds, according to the gamma-ray comparison with the 1911 radium standards, to 24.23 mg of radium element, as of June 1934. Information on all four national standards is summarized in table 1.

TABLE 1. *Description of four national radium standards*

	A[1] U. S. primary radium standard	*B* British primary radium standard	*C* Canadian national radium standard	*D* U. S. primary radium standard
Reference numbers	5437, XIV	5432	XIII	5440, XV
Radium content as given by:				
1. *Hönigschmid's* weighings.	38.23 mg	15.60 mg	.....	20.45 mg
2. Comparison with Paris and Vienna 1911 standards, as of June 1934	38.13	.....	24.23 mg	20.38 mg
Length of glass tube	36 mm	38.8 mm	10.5 mm	37 mm
Internal diameter of tube	3 mm	3 mm	1.5 mm	3 mm
Tube wall thickness	0.27 mm	0.27 mm	0.25 mm	0.27 mm

[1] For convenience, *A*, *B*, *C*, and *D* are used here and elsewhere in this paper to identify these radium standards.

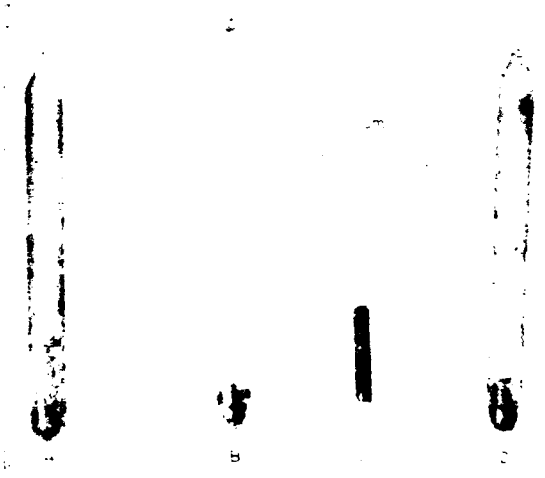


FIGURE 1. *Four national radium standards.*

A, American; *B*, British; *C*, Canadian; and *D*, American.

In view of the uncertainties that exist and the differences between the Hönigschmid weights and their weights as derived by comparison with the 1911 Paris and Vienna standards [3], it has recently been suggested that new radium standards be prepared from about 1 g remaining of Hönigschmid's original "atomic-weight" material. Another possibility lay in a recheck of the present standards. With this end in view, United States primary radium standard 5440, XV was taken to the United Kingdom in the summer of 1952 and to Canada in the autumn of the same year. At the National Physical Laboratory, in Teddington, and at the National Research Council Laboratories, in Ottawa, it was compared, by gamma radiation, with the British primary radium standard, 5432 [4] and the Canadian national radium standard, XIII [5]. The results obtained by these laboratories are discussed later in connection with the data given in table 3.

The question also arises as to what is desired in a radium standard. In order to derive the mass of any radium preparation in terms of the standard by gamma-ray measurements it is necessary to know both the absorption of the containers of the preparation and standard and also the self-absorption of the radium salts themselves. In NBS certificates the results are stated in terms of milligrams of radium when contained in a Thüringen glass tube having a wall thickness of 0.27 mm, together with an empirical absorption correction for the container in question. Only calorimetric measurements can give the ratios of the true radium contents, irrespective of absorption but in this case it is necessary to know the date of sealing of the preparation in order that correction may be made for the growth of polonium. A small fraction of the gamma-ray energy is absorbed and measured by the calorimeter, but any difference in absorption between two sources will represent only a small correction to the already small contribution of gamma-ray energy emission (about 7%) to the total energy emission.

3. Measurements With the NBS Standard Electroscope

The NBS standard electroscope [6] and measuring system were used, without modification, for this comparison of four national radium standards. The ionization chamber consists of a 10-cm cube free-air volume, with walls made of 1 cm of lead and a ½-cm aluminum inner lining. A gold leaf is suspended near the center of the chamber. A 10-μ quartz fiber at the free end of the leaf provides a fine line for projection. The fiber image is magnified approximately 100 times and projected onto a metric scale. The discharge of the electroscope is measured by timing the transit of the image between two fixed points on the scale 6 cm apart.

The source indexing system consists of a V-shaped trough of 3/32-in. Lucite on an aluminum stand. The stand can be moved along a line perpendicular to the face of the ionization chamber or rotated about its own vertical axis. Preparations are centered in the trough opposite the center of the chamber, so that measurements are made perpendicular to the axes of symmetry of the preparations.

The four standards were measured relative to each other by comparison of each of the six possible combinations of pairs. Independent measurements were made on each pair by each of three different observers at source distances of 66.5 cm and 74.1 cm from the chamber. The entire series of measurements was repeated twice.

The following procedure was adopted for comparing each pair of standards:

1. The trough was placed at the distance selected and parallel to the chamber face.
2. A standard was held horizontally and tapped lightly until the salt was distributed uniformly along the length of the capsule, as in figure 2.
3. The standard was placed in the trough and centered.
4. Three observations of the discharge time were made and recorded.
5. The trough was rotated 180 degrees, and three more observations were made.
6. Procedures 1 to 5 were repeated with the second standard of the pair.
7. Procedures 1 to 6 were repeated for both members of the pair at the second distance from the electroscope.

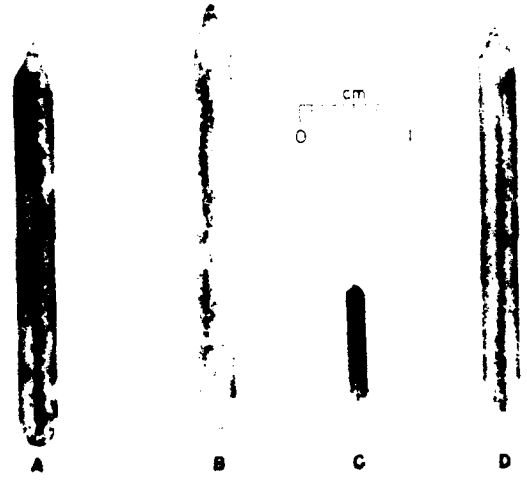


FIGURE 2. *Four national radium standards, with the grains of salt in the three Hönigschmid standards distributed along the length of the tubes.*
A, American; *B*, British; *C*, Canadian, and *D*, American.

4. Comparison by Geiger-Müller Counter

The Geiger-Müller counter used for this comparison was a neon-halogen-filled tube. The tube itself was surrounded by a sheath of lead ¾ inch thick so that the soft gamma rays, the spectrum of which might be varied by source absorption to a greater extent than that of the higher-energy gamma rays, would not be counted. The resolving time of the counter was determined by the two-source method to be 211 μsec ± 5 percent. The correction for re-

solving time applied to the data ranged from 1.1 to 2.7 percent.

The source holder of the NBS standard electroscope was used to position each standard in turn in these measurements, and the standards were tapped so that, in the case of the more loosely packed Hönigschmid standards, the grains would be distributed uniformly along the tube.

In order to eliminate any possible effects due to drift a series of measurements was carried out on each pair of international standards. Thus, in the comparison of A and B, measurements were carried out with A and B arranged in "packages" in the following order: $A, B, B, A; B, A, A, B; A, B, B, A;$ and finally, B, A, A, B. Similar package measurements were made on each of the other five pairings of the four international standards.

A total of about 80,000 counts was taken on each of the 16 members of the 4 packages comprising a pair comparison. Thus in the comparison of A and B a total of some 640,000 counts were made with A in position and 640,000 with B.

5. Comparison by Scintillation Counter

The scintillation counter consisted of a thallium-activated sodium-iodide crystal mounted on the face of a photomultiplier tube. The resolving time of the counter and amplifier was 5 μsec ± 10 percent, and the corrections applied to the data varied from 0.3 to 0.8 percent. The discriminator was set to accept pulses corresponding to gamma-ray energies greater than 1 Mev. Thus, as for the NBS standard electroscope and the Geiger-Müller counter, the effect of source self-absorption of the lower-energy gamma rays should not be apparent. The sodium-iodide crystal and photomultiplier were mounted adjacent to the Geiger-Müller counter so that counts on each source could proceed concurrently with both counting systems. Exactly the same pairing and packaging order of sources as was used for the Geiger-Müller counter comparison was, ipso facto, also used in the scintillation-counter measurements. The counts for each source in position were of the order of 400,000 compared with 80,000 in the case of the Geiger-Müller counter.

6. Measurements With the Radiation Balance

A modification of the radio-balance originally designed by Callendar [7] for the measurement primarily of radiant energy has recently been described [8], which is suitable for the measurement of the energy emission from radioactive materials. This modification of the radio-balance has been renamed the radiation balance, its most important feature being the ability to balance the energy emission from a radioactive source either against Peltier cooling or the energy emission from another radioactive source, or both.

None of the radiation balances constructed previously was large enough to accommodate the large Hönigschmid standards, and, accordingly, a new one was constructed for this purpose. This balance is described in detail separately in this issue [9]. It differed essentially from the first one, however, in that its larger cups were made from gold instead of copper.

7. Radiations Measured

The radiation from radium in equilibrium with all its products consists of five energetic alpha-particle groups, including that of polonium; three main groups of beta particles, the most energetic being that from the transition of radium E to radium F with a maximum energy of 1.17 Mev; and a complexity of gamma rays, the most energetic being from the excited levels of radium C'.

Three of the methods described here and used to compare the radium contents of the four national radium standards were essentially gamma-ray comparisons. With the thicknesses of lead used, or the discriminator setting, the chief contribution to the gamma-ray effect would be from the energetic radium C' gamma rays (above 0.6 Mev in the case of the electroscope and Geiger-Müller counter and above 1 Mev in the case of the scintillation counter).

In contrast, the radiation balance measures primarily the energy emitted in corpuscular form. Some 93 percent of the energy produced by radium and its daughters down to radium D is associated with particulate emission, the remaining 7 percent of the energy produced being associated with the gamma radiation. The wall thickness of the gold cups was such as to absorb completely the most energetic beta particles from radium E. Some 12 percent of the energy associated with the gamma rays is also absorbed. Of the 7 percent of the total energy produced that is associated with the gamma-ray emission, another 1 percent (for the Canadian standard) or 1.5 percent (for the Hönigschmid standards), corresponding, respectively, to 0.07 and 0.1 percent of the total energy produced, will be absorbed in the sources themselves. The difference of 0.5 percent between the source self-absorption of the Canadian and Hönigschmid standards corresponds therefore to a difference of only 0.035 percent of the total energy produced, which is negligible. Any smaller differences in gamma-ray source self-absorption of the three Hönigschmid standards are also therefore negligible so far as the measurements in the radiation balance are concerned. The alpha-particle and beta-particle absorption is complete; a correction must be made, however, for the growth of radium E and polonium, which will not be in equilibrium with the radium.

8. Results

The results obtained with the radiation balance, measuring the sources singly and in every combination of pairs, are summarized in table 2. In this table the order of measurement is represented by reading from left to right and down the table.

From the results in table 2 the following best estimates for the energy absorbed (in microwatts) from sources A, B, C, and D have been deduced:

A	B	C	D
6293.4	2569.8	4131.0	3360.7

TABLE 2. *Energy absorption, in microwatts*

Source	Energy absorbed	Source	Energy absorbed	Source	Energy absorbed	Source	Energy absorbed
B	2571.0	A-B	3727.1	D-B	788.6	C-B	1561.2
C	4127.1	C-D	776.0				
D	3371.6	A-D	2935.6				
A	6285.2	A-C	2164.2				

In table 3 are shown the complete results for the six pairs of standards, using the NBS standard electroscope, Geiger-Müller counter, scintillation counter, and radiation balance. In the last line of the table are shown the weight ratios for the same six pairs. The weight of the Canadian standard (*C*) is, however, only a derived weight, and for this reason, any ratio involving this derived weight is shown in quotation marks. The ratios *A*/*B*, *A*/*D*, and *B*/*D* are, however, the ratios of Hönigschmid's own weighings.

TABLE 3. *Adjusted results for the ratios of the four international standards*

Method	*A*/*B*	*A*/*C*	*A*/*D*	*B*/*C*	*B*/*D*	*C*/*D*
NBS standard electroscope	2.441	1.570	1.870	0.6429	0.7661	1.192
Geiger-Müller counter	2.461	1.582	1.885	.6430	.7659	1.191
Scintillation counter	2.478	1.579	1.889	.6370	.7624	1.197
Radiation Balance	2.449	1.523	1.873	.6220	.7647	1.229
Weighing	2.451	"1.578"	1.869	"0.6438"	.7628	"1.185"

For comparison with these values the ratios obtained by Perry [4], using the NPL standard ionization chamber with gold-leaf electroscope, an ionization chamber with a Lindemann electrometer, and a Geiger-Müller counter for *B*/*D* were, respectively, 0.7669, 0.7657, and 0.7669. The result obtained for the gamma-ray ratio *C*/*D* by Michel [5], using the NRC precision ion chamber and Lindemann electrometer was 1.192. Michel, from geometrical considerations, then calculated the source absorption of each standard and corrected the gamma-ray ratio to give a weight, or content, ratio of *C*/*D* equal to 1.185. The direct gamma-ray ratios obtained both by Perry and Michel are in excellent agreement with the results shown in table 3.

A check on the internal consistency of the results shown in table 3 can be provided by assuming that *A*, *B*, and *D* are so much alike that there are negligible differences in source absorption for high-energy gamma rays, and none at all in the case of the calorimeter, where 93 percent of the energy absorbed is particulate, so that any change due to absorption of the 7 percent of gamma rays and secondary electrons would be even more negligible. A check can then be run on the results for *A*, *B*, and *D* by dividing the quantity characteristic of each in each determination by the Hönigschmid weight of each standard. This characteristic quantity is scale divisions per second for the NBS standard electroscope, counts per second for the counters, and microwatts for the radiation balance. In each case the characteristic quantity is as of February 1954, and the mass of radium element is as of June 1934. It is not necessary for this check to correct for the 20-year decay of radium as this is the same constant for each standard.

The results of this internal precision check are shown in table 4, in which the figures quoted are the characteristic quantity, divisions, or counts per second or microwatts, divided by the mass of radium element present and normalized to make the "best average" equal to 100.00 in each case. This best average is obtained by dividing the sum of all the three radioactive effects by the sum of all three masses.

TABLE 4. *Radioactive effect per milligram of radium element*

(Normalized to make the best average equal to 100.00)

Method	*A*	*B*	*D*	Best average	Standard deviation of the individual results
NBS standard electroscope	99.93	100.31	99.88	100.00	0.23
Geiger-Müller counter	100.31	99.90	99.51	100.00	.40
Scintillation counter	100.52	99.42	99.47	100.00	.62
Radiation balance	100.03	100.10	99.86	100.00	.13

The values of the best average should, in turn, enable one to form an estimate of the precision of Hönigschmid's weight determinations, in which, according to Hönigschmid himself [2], the error of a single weighing was not more than 0.02 mg. A statistical survey of the results was carried out with the cooperation of W. S. Connor and W. J. Youden, and resulted in the best estimates of the mass of radium element in *A*, *B*, and *D* given in table 5. The methods adopted to arrive at these best estimates, and also the best estimates given in table 3 for the ratios of pairs of standards, are described by Connor and Youden in part 2 of this paper.

TABLE 5. *Best estimates, in milligrams, of the masses of the Hönigschmid radium standards, as of June 2, 1934*

Standard	*A*	*B*	*D*
Hönigschmid's mass	38.23	15.60	20.45
Mass derived from NBS standard electroscope	38.227	15.611	20.446
Mass derived from Geiger-Müller counter	38.235	15.598	20.443
Mass derived from scintillation counter	38.235	15.595	20.444
Mass derived from the radiation balance	38.235	15.608	20.435

9. Mass of Radium Element in the Canadian National Standard

By comparing the calorimetric ratios given in table 3 with the "weight" ratios, it is clear that the derived weight of the Canadian national radium standard (*C*) is low by about 3 percent. However, this does not allow for the difference in sealing date,

which involves a compensating polonium-growth correction of about 1.8 percent. By comparison of the "weight" ratios with the NBS standard electroscope ratios, it is also confirmed that no source self-absorption correction could have been made in deriving the certified weight of radium in the Canadian standard. However, from the data available it is possible to derive a value for this mass of radium.

The experimentally determined ratios of the energy absorbed in the radiation-balance cups per unit mass of radium element for *A*, *B*, and *D* are 164.62, 164.73, and 164.34 μw/mg, respectively. Taking the best average value of 164.58 μw/mg of radium element, the mass of the radium in the Canadian national radium standard is found to be equal to 25.10 mg as of June 1934, uncorrected for the growth of polonium or of radium E.

Using the Curie-Yovanovitch equation, as corrected for new values of the decay constants by Jordan [8, 10], the energy increments due to growth of polonium-210 in *A*, *B*, and *D*, on the one hand, and in *C*, on the other, are found to be equal, respectively, to 12.2 and 16.2 cal g^{-1} hr^{-1} inclusive, of nuclear recoil energy, the separation and sealing dates being, respectively, May 25, 1934, and June 2, 1934, for the Hönigschmid standards, and June 1924 and June 1930 for the Canadian national standard. The growth of radium E will contribute, in proportion, another 0.8 and 1.0 cal g^{-1} hr^{-1}. Subtracting the contributions of polonium-210 and nuclear recoils and of radium E from the energy absorbed from *A*, *B*, and *D* in the radiation-balance cups gives a total energy absorption for all three sources equal to 11103.0 instead of 12223.9 μw (as of February 1954).

In the case of the Canadian national standard, an energy production of 17.2 cal g^{-1} hr^{-1} by polonium-210 and radium E corresponds to 20.0 μw/mg of radium element, which, by a second approximation, is found to be equivalent to 489.7 μw/24.48$_5$ mg of radium element (the mass of radium as of June 1924). The corrected energy absorption from the Canadian national radium standard is therefore 3641.3 instead of 4131.0 μw, as of February 1954. The radium content of the Canadian national radium standard, as of June 1934, is then obtained by multiplying the total weight of the Hönigschmid standards (76.28 mg as of June 1934) by the ratio of the corrected energy absorptions of February 1954. This gives the result that there were 24.36 mg of radium element in the Canadian national standard, as of June 1934. This value will, if anything, be on the low side, however, as some radium D on the walls of the original two tubes may have been lost on transfer when the Canadian standard was resealed in June 1930. In this event, the polonium-210 correction will have been too great.

10. Summary of Results

As a result of this intercomparison of national radium standards, the ratios of the weights ascribed to three of them by Hönigschmid have been confirmed. It would appear that the weights derived from the comparison of the two United States standards with the 1911 Paris and Vienna standards are, therefore, too low; unless it were assumed that all of Hönigschmid's mass determinations were low in the same ratio. However, this is to be discounted because the Berlin standard was, by comparison with the 1911 standards, found to have a greater weight than that determined by Hönigschmid [1].

Relative to the Hönigschmid weights, the Canadian national radium standard is found to have a mass of radium element equal to 24.36 mg, which indicates that no correction for difference in source self-absorption was made in its comparison with the 1911 Paris and Vienna standards. The difference between this value and that obtained by comparison with the 1911 Paris and Vienna standards (24.23 mg as of June 1934) would indicate a self-absorption correction of 0.53 percent. The absorption correction determined by Michel [4] was 0.94 percent; the difference between these two values could be a measure of the loss of radium D and polonium-210 in the transfer of June 1930.

Most grateful acknowledgments are made to the following: The Director of the National Physical Laboratory and the President of the National Research Council, for the loan of the British primary radium standard and the Canadian national standard; to W. E. Perry and W. S. Michel, respectively, for their helpful cooperation and for the transportation to Washington, D. C., of these standards; to W. J. Youden, for many helpful and most valuable discussions on the planning of the experiments, and to him and W. S. Connor, Jr., for discussion of the final results; to H. H. Seliger for providing the Geiger-Müller and scintillation counter equipment and for advice on its operation; and to L. F. Paolella for valuable assistance in carrying out the readings on the NBS standard electroscope and the Geiger-Müller and scintillation counters.

11. References

[1] C. Chamié, Sur la nouvel étalon international de radium, J. phys. radium [8] **1,** 319 (1940).
[2] O. Hönigschmid, Geschichte und Herstellung der primären Radium-Standards, Anz. Akad. Wiss. Wien **82,** 30 (1945).
[3] Stefan Meyer, Über die Radium-Standard-Präparate, Anz. Akad. Wiss. Wien **82,** 25 (1945).
[4] W. E. Perry, A gamma-ray comparison of British and United States national radium standards, Proc. Phys. Soc. (in press).
[5] W. S. Michel, The intercomparison at the National Research Laboratories of a primary radium standard of the National Bureau of Standards and the Canadian national radium standard, National Research Council of Canada Report No. PR–192 (June 23, 1953).
[6] L. F. Curtiss, A projection electroscope for standardizing radium preparations, Rev. Sci. Instr. **16,** 363 (1928).
[7] H. L. Callendar, The radio-balance. A thermoelectric balance for the absolute measurement of radiation, with applications to radium and its emanation, Proc. Phys. Soc. **23,** 1 (1911).
[8] W. B. Mann, Use of Callendar's "radio-balance" for the measurement of the energy emission from radioactive sources, J. Research NBS **52,** 177 (1954) RP2486.
[9] W. B. Mann, A radiation balance for the microcalorimetric comparison of four national radium standards, J. Research NBS **53,** 277 (1954) RP2545.
[10] K. C. Jordan (private communication).

Part 2. Statistical Procedures and Survey

W. S. Connor and W. J. Youden

The statistical analysis of the observations on the four national radium standards is discussed. The readings made with the electroscope, Geiger-Müller counter, and scintillation counter were adjusted by one formula, and the readings made with the radiation balance by a different formula. In each case the adjusted values of the standards satisfy a consistency criterion. Finally, the adjusted values were improved by making use of the proportional relationship between the masses and the radioactive effects of the standards.

1. Introduction

Four national radium standards were recently compared at the National Bureau of Standards, as described in part 1 of this paper. The unusual opportunity associated with the presence of four standards in one laboratory directed attention to certain statistical aspects of the intercomparison. The experimental procedures and results are described in part 1. Part 2 discusses the statistical analysis.

When two standards are compared, careful measurements provide an estimate for the value of one standard in terms of the other. A standard error may be calculated for this estimate. A third standard makes possible the additional experimental evaluation of each of the first two standards in terms of the third.

Suppose that three standards A, B, and C are available. The experimental ratios a/b, b/c, c/a may each be determined by using exactly the procedure that would have been employed if just two standards had been available. None of the measurements made on A in estimating a/b are used in the estimation of c/a. Additional data for A are taken to determine c/a. There is a considerable advantage in this method because the precision of the comparison is improved by alternating the readings on the two standards under comparison. This alternation reduces the effects of drift in the instruments and changes in the environment. As soon as the ratios a/b, b/c, c/a have been determined there is a simple test for the consistency of the three ratios. The product of the three ratios should be unity. The discrepancy between this product and unity provides a measure of the errors in these ratios.

A similar consistency criterion was applied to the six ratios determined by the electroscope, Geiger-Müller counter, and scintillation counter. Because a different statistical treatment was required for the measurements made with the radiation balance, those measurements are discussed separately.

The last section describes how the masses of the standards were used further to improve the estimates of the standards.

2. Comparison of the Standards by Means of Electroscope, Geiger-Müller Counter, and Scintillation Counter

Using these methods, environmental conditions common to paired measurements introduce a common multiplicative error in the measurements. It is advantageous to express the results of paired measurements as ratios to eliminate this error.

There were four standards, A, B, C, and D. Therefore, the following six ratios could be determined experimentally:

$$a/b \quad a/c \quad a/d \quad b/c \quad b/d \quad c/d.$$

These provide opportunities to test the consistency of the data. For example, the products

$$a/b \times b/c \times c/a$$

$$a/b \times b/d \times d/a$$

$$a/c \times c/d \times d/a$$

$$b/c \times c/d \times d/b$$

should all be equal to unity. The discrepancies between these products and 1.0000 reveal the errors of the measurements. It is proper to make use of the information that the products should be exactly equal to one. The measured ratios may be adjusted by a least-squares technique to obtain new ratios $\hat{A}/\hat{B}$, $\hat{A}/\hat{C}$, etc., which do in fact multiply out to unity for all combinations that should give unity. This includes not only three factor combinations such as

$$\hat{A}/\hat{B} \times \hat{B}/\hat{C} \times \hat{C}/\hat{A}$$

but also four factor products

$$\hat{A}/\hat{B} \times \hat{B}/\hat{C} \times \hat{C}/\hat{D} \times \hat{D}/\hat{A}.$$

The adjustment formula used on the data shown in table I is of the form

$$\frac{\hat{A}}{\hat{B}} = \sqrt[4]{\left(\frac{a}{b}\right)^2 \left(\frac{a}{c} \times \frac{c}{b}\right)\left(\frac{a}{b} \times \frac{d}{b}\right)},$$

where the lower case letters indicate the measured ratios.[1] The adjusted values (see table 3 in part 1)

[1] This adjustment formula is related to the adjustment formula for the difference between the estimates of two treatment effects in a balanced incomplete block (BIB) design, see R. L. Anderson and T. A. Bancroft, Statistical theory in research, p. 252 (McGraw-Hill Book Co., Inc., New York, N. Y., 1952). Since the two measurements in a pair, as a and b or c and d, are subject to a common multiplicative error, the logarithms of the two measurements in the pair are subject to a common additive error. Hence, the BIB design formula applies for the difference between the logarithms of the adjusted values, as $\log \hat{A} - \log \hat{B}$, and by taking antilogarithms, the above formula is obtained.

have the property that

$$\frac{\hat{A}}{\hat{B}}=\frac{\hat{A}}{\hat{C}}\times\frac{\hat{C}}{\hat{B}}=\frac{\hat{A}}{\hat{D}}\times\frac{\hat{D}}{\hat{B}}.$$

The observed values do not meet this consistency requirement. The reconciliation among the results effected by the above least-squares technique introduces each standard symmetrically in the computation pattern and does not single out any one standard as a superstandard. After the relative values have been established, one standard may be given an agreed value, whereupon all other standards are determined without changing the relative values.

TABLE 1. *Experimental results for the ratios of four standards*

Method	a/b	a/c	a/d	b/c	b/d	c/d
Electroscope	2.4438	1.5675	1.8703	0.64246	0.76650	1.1918
Geiger-Müller counter	2.4746	1.5785	1.8784	.64489	.76789	1.1920
Scintillation	2.4847	1.5710	1.8930	.63921	.76186	1.1953

The above least-squares adjustment has long been used for other comparisons. Recently, it has been found that certain subsets of pairs selected from all possible pairs lead to convenient least-squares estimates.[2] Given that a reasonably small number of pairs will suffice to interrelate all the standards, there would appear to be some chance of success for an international program of comparison. Once a properly selected subset of pairings was obtained, the various national standards could be tied together with values that would give consistent comparisons among the standards.

3. Radiation-Balance Measurements

The radiation balance used in this work was suitable for measuring either a proportion of the energy emitted by one standard or the same proportion of the difference in energies emitted from two standards. This difference is determined by one measurement. The schedule of measurements included separate measurement on the four standards as well as the six possible differences between them. The precision of measurement of a difference was the same as the precision of measurement of a single standard.

Typical formulas for the least-squares estimates [3] for the 10 quantities follow:

$$\hat{A}=\tfrac{2}{5}a+\tfrac{1}{5}[(a-b)+b]+\tfrac{1}{5}[(a-c)+c]+\tfrac{1}{5}[(a-d)+d]$$

$$(\hat{A}-\hat{B})=\tfrac{2}{5}(a-b)+\tfrac{1}{5}[(a-c)+(c-b)]+\tfrac{1}{5}[(a-d)+(d-b)]+\tfrac{1}{5}a-\tfrac{1}{5}b.$$

The quantities a, b, $(a-b)$, etc., are measured quantities. The value for $(\hat{A}-\hat{B})$ given by the above formula will agree exactly with the result obtained by subtracting the adjusted estimate $\hat{B}$ from the adjusted estimate $\hat{A}$. This was not true for the recorded values. The total amount of energy measured for the standards is left unaltered by the adjustment. Slight shifts take place in a, b, c, d, $(a-b)$, etc., to achieve consistency among the results. The discrepancies between the measured quantities and the corresponding adjusted values afford a measure of the precision of the measurements. The calculation is shown in table 2. It should be noted that no quantity was measured twice. The replication is concealed. There are, of course, only four standards; that is, four quantities to be determined from the ten observations. This leaves six contrasts, i. e., six degrees of freedom, available for estimating the standard deviation.

TABLE 2. *Calculation of standard deviation, in microwatts, for radiation balance*

Standard	Observed	Adjusted	Difference	(Difference)^2
	μw	μw	μw	μw
A	6285.2	6293.4	8.2	67.24
B	2571.0	2569.8	1.2	1.44
C	4127.1	4131.0	3.9	15.21
D	3371.6	3360.7	10.9	118.81
$A-B$	3727.1	3723.6	3.5	12.25
$A-C$	2164.2	2162.4	1.8	3.24
$A-D$	2935.6	2932.7	2.9	8.41
$C-B$	1561.2	1561.2	0.0	0.00
$C-D$	776.0	770.3	5.7	32.49
$D-B$	788.6	790.9	2.3	5.29

$$\text{Standard deviation}=\sqrt{\frac{264.38}{6}}=6.6\mu w.$$

4. Masses of the Radium Standards

Standards A, B, and D were made from the same supply of radium salt. The weighings were made in the same day by Hönigschmid and are considered to have a maximum error of 0.02 mg. The various properties of the three Hönigschmid standards measured by the several methods used in this intercomparison are believed to be directly proportional to the masses of the standards. All the methods give relative values for the standards. In addition, the radiation balance measures the difference between any two standards directly. Standard D was arbitrarily given the value of unity and the values for A and B expressed relative to it. Table 3 contains some of the adjusted ratios from table 3 of part 1, including the ratios derived from Hönigschmid's weighings.

[2] W. J. Youden and W. S. Connor, Making one measurement do the work of two, Chem. Eng. Progr. 49, 549 (1953); and W. J. Youden and W. S. Connor, New experimental designs for paired observations, J. Research NBS 53, (1954) RP2532.

[3] For a discussion of the method of least squares, see R. L. Anderson and T. A. Bancroft, Statistical theory in research, p. 155 (McGraw-Hill Book Co., Inc., New York, N. Y., 1952).

TABLE 3. *Value of standard, when D equals 1.000*

Method	A	B	D
Scintillation counter	1. 889	0. 7624	1. 000
Geiger-Müller counter	1. 885	. 7659	1. 000
NBS standard electroscope	1. 870	. 7661	1. 000
Radiation balance	1. 873	. 7647	1. 000
Weighing	1. 869	. 7628	1. 000

For each method of measurement a plot may be made of the values of the standards against the corresponding masses. The resulting points should lie along a straight line that passes through the origin. Let m denote the mass and r the radioactive effect given by any one method of measurement in the relative units of table 3. If the error in r is k times as large as the error in m (as measured by the standard deviations), the slope b may be computed from the quadratic

$$b^2+\frac{k^2\Sigma m_i^2-\Sigma r_i^2}{\Sigma m_i r_i}b-k^2=0,$$

where m_i and r_i ($i=A, B, D$) are the masses and the corresponding radioactive effects for standards A, B, and D from table 3.[4] This method of determining b has the property that the sum of the squares of the perpendicular distances of the points (km_i,r_i) from the line $r=b(km)$ is minimized.

[4] For a discussion of this method, see W. Edwards Deming, Statistical adjustment of data, Exercise 6, 184 (John Wiley & Sons, Inc., New York, N. Y., Nov. 1944).

In general, any particular plotted point will not be located exactly on the fitted line. The plotted points are subject to errors of observation. The "best" estimates of the coordinates for the point are taken to be the coordinates of the point on the line nearest to the plotted point. These coordinates, m_i' and r_i', are

$$m_i'=\frac{m_i+br_i}{1+b^2} \quad \text{and} \quad r_i'=b\frac{m_i+br_i}{1+b^2}.$$

This procedure for fitting lines was followed for each of the lines relating the measured radioactive property to the mass. For the electroscope, Geiger-Müller counter, scintillation counter, and radiation balance, k was taken as 2, 4, 5, and 1, respectively. These values correspond to the errors given by table 4 of part 1, except for the scintillation counter, for which 5 was used instead of 6.

To obtain estimates for the masses, each value of m_i' was multiplied by Hönigschmid's value for D, i. e., 20.45 mg. These estimates are recorded in table 5 of part 1. In every case the result agreed with the assigned mass within the claimed weighing error. It is particularly interesting to observe that the estimates obtained from the line, using the radiation balance results, confirm the assigned masses. For this line the errors in m and r were taken to be the same, and therefore any displacement of the point to bring it on the line required equal changes in the experimental values for mass and energy.

WASHINGTON, May 27, 1954.

PHYSICAL MEASUREMENTS AND EXPERIMENT DESIGN

W. J. YOUDEN

National Bureau of Standard Washington, D. C.

ABSTRACT

Each field of experimental inquiry poses problems that are characteristic of the field. Research programs are, in general, planned to take advantage of the special features that broad classes of problems present. One broad class of problems deals with the determination of physical constants and the calibration of instruments. This paper discusses the statistical aspects of physical measurements and suggests some experimental programs that may be useful to those concerned with the determination of physical constants and with calibration procedures.

CONTENTS

INTRODUCTION

The remarkable success of experimental design in agricultural field trials was aided by the spectacular diminishment of the experimental error that came about through grouping the plots with different treatments into compact blocks. Replication and randomization insured the calidity of the estimate of error and made possible the unambiguous interpretation of the data. Agricultural experiments prompted statisticians to devise various way of grouping the experimental plots into blocks. The extension of statistical design into chemical research stimulated the further development of new designs. It appears that every major field of research has problems which invite the invention of new designs or the adaptation of old designs. This paper examines the special opportunities for experiment design in measurements of physical properties.

EXPERIMENTATION IN THE LABORATORY

Agricultural field trials were characterized by a number features such as :

a) Large experimental errors

b) Extensive replication

c) Ease of randomization

d) Considerable freedom in the number of plots per block

e) Frequently a large number of experimental items

f) All the data obtained at once at the end of the experimental period

g) Interactions between factors

h) Interest centered on comparisons, not absolute values.

In contrast to the above, experimentation in the laboratory brings a controlled environment, much smaller errors, often very little replication. The land blocks of the agricultural experimenter usually became identified with instruments, or days, or operators. The plots may become the "heads" or different positions on a test machine. One run with the machine may constitute a block, and thus the size of the block is determined by the structure of the machine. Usually the number of items under comparison is smaller than in field trials and sometimes randomization is an expensive or difficult condition to meet. More important, the data are obtained sequentially so that the experimenter may examine the results of the last run before beginning the next test. This sequential process of gathering the data generally makes the experimenter unwilling to commit himself to a large rigid program of work.

THE MEASUREMENT OF PHYSICAL QUANTITIES

Physical measurements are continually being made to improve the accuracy of important physical constants. Periodically there are repeat determinations of the gravitation constant, g, the velocity of light, the astronomical unit of distance (distance from the earth to the sun), and many other fundamental constants of nature. Usually the results of these investigations lead physicists to regard statistics as unable to make a worthwhile contribution to their problems. Whatever the physical constant the story is the same. A careful study is made in a given laboratory using the utmost care to construct an assembly of equipment for making the measurements. A considerable number of repeat measurements are made and from these data an average and a standard deviation is obtained. When the results from different laboratories are compared, the differences among them are invariably very much greater than would be expected on the basis of the estimates of the standard deviations.

Physicists correctly concluded that these standard deviations measured only the local precisions and threw no light whatever on the presence of systematic errors associated with particular assemblies of apparatus. The investigators were concerned with absolute values and statistical designs were used for comparisons. The efforts of investigators were therefore concentrated upon greater care in the calibration of equipment and upon ingenious arrangements to compensate automatically for some of the possible sources of errors.

The very notion of a physical constant carries with it the implication that the value of the constant should be independent of the particular assembly of equipment used to determine the constant. Of course, there are many components of the equipment whose properties such as diameters of orifices, lengths and resistances, together with operating conditions such as temperatures, pressures and voltages which have an influence on the observed result. The essence of the matter is, that if these various properties and operating conditions are known and entered into the proper formulas, the outcome should be the value desired. It is traditional in careful work to vary the operating conditions one at a time and to collect evidence that, when due allowance is made for the change, the determinations made before and after the change show acceptable agreement. Less often actual substitution of a component of the apparatus will be made. One resistance coil may be replaced by another. It will not matter that there is a difference between the two resistances. What does matter is that the resistance of each coil be known so that determinations using first one coil and then the other coil will show acceptable agreement.

The experimental problem just discussed in connection with the determination of fundamental physical constants also arises in the evaluation of the properties of substances. Density, viscosity, boiling points, conductivity ; the list of properties is very long. A great amount of effort goes into the revision of old values and into the determination of properties for the unending production of new substances. The preparation

of reference samples with stated properties and the calibration of instruments both involve the use of apparatus and procedures. The correction to be applied at a particular scale point on an electrical instrument is a quantity that the calibrating laboratory undertakes to establish. The calibration laboratory must satisfy itself that the correction reported is only to a small degree influenced by the particular equipment and technique used in the calibration process.

STATISTICAL CHARACTERISTICS OF PHYSICAL MEASUREMENTS

If a laboratory does have the means to put together different assemblies of equipment and to vary some of the operating conditions, the collection of results obtained correspond to what the agricultural experimenter terms a "uniformity trial". Sometimes all the plots in a large area are given the same treatment. The observations from these plots should agree except for the experimental errors. In the same way, if the substitution of components has been without effect and any deliberate changes in the operating conditions properly allowed for, the experimental results should show only the variation of random errors. Ideally there should be no greater variation among such results than in a series of repeat measurements where no changes in apparatus or operating conditions are introduced. The precision error is therefore the appropriate criterion for judging whether or not deliberately introduced changes do have an effect.

There are always enough repeat measurements to furnish a good estimate of the standard deviation. The skill of the experimenter insures that any "effects" that are associated with substitutions of components or other changes are of the order of magnitude of the precision standard deviation. If the investigator is to have a reasonable chance of detecting an effect equal to sigma, he will need to make around 15 repeat measurements before after making the change. Such checking of a number of aspects of the equipment and procedure soon multiplies the number of measurements especially under the traditional procedure of changing just one item at a time. Actually it is not enough to reduce such individual effects due to substitutions to about the magnitude of σ. The signs of these effects may be either positive or negative. The observed result is the net *sum* of such effects and this is undoubtedly one of the main reasons for the disagreement between reports from different laboratories. Consequently there is an acute need for more efficient experimental designs than the "change one factor at a time" procedure.

There is one important encouraging element in these experimental programs. The possible effects that are under study can be taken as purely additive. In other words, there is no reason to fear the presence of interactions between the factors. The simplification in experiment design that results is so marked that it is necessary to indicate the argument for the absence of interactions. Condider two similar rods calibrated for length, either of which may be used in the apparatus for making the measurement. If no other factor is changed, a series of mea-

surements with each rod may reveal a small effect associated with the substitution. Probably the calibrations of the rods are slightly in error. Now suppose the two series of measurements are repeated with the apparatus at a slightly higher temperature than was maintained during the first trials with the rods. The higher temperature causes the rods to expand in length and the experimenter allows for this expansion in his calculations. The difference in results with the two rods, as revealed in the first trials, will be very small, possibly near the limit of detection. Certainly the temperature change has a substantial effect on the *lengths* of the rods but the *change in the difference between results* with the two rods will be completely undetectable. The whole difference between results with the two rods is difficult enough to establish. Consequently we may take the difference in results with the two rods to be independent of the temperature.

Experimenters supply ample evidence of the fact that interactions can be ignored. Work starts with some initial assembly of equipment and specified conditions. Subsequently changes are made, one by one, to ascertain whether or not the measurement is acceptably immune to such changes. The experimenter has no misgivings at all that he will overlook some effect solely because of the particular assembly and conditions that happen to constitute the reference set. If the effects did depend on the choice of the reference set, the research would be vastly more complicated, the very concept of a physical constant would lose sharpness, and tables of critical constants would have to specify the experimental details.

REQUIREMENTS FOR EXPERIMENT DESIGNS FOR PHYSICAL MEASUREMENTS

The discussion thus far has covered in some detail the special characteristics of experimentation directed to the evaluation of physical constants. Statisticians can propose experimental designs that combine high efficiency in the detection of very small effects together with a satisfactory estimate of the physical constant. It is the latter requirement of an unbiased estimate of the mean that has special importance.

Long ago Yates [1] pointed out that a 1/16 fraction of a 2^7 factorial provided mutually orthogonal estimates of the seven main effects. The fact that these estimates were confounded with interactions of the factors led Yates to warn his readers that it would rarely be wise to assume the absence of interactions. The point has been made above that, in physical measurements, the interactions, if present at all, are of negligible magnitude in comparison with the main effects. Table 1 shows this design where zero and one denote the alternative choices for each of the seven factors A, B, C, D, E, F and G.

Table 1. Seven factors with two choices designated by zero and one.

Combination	Factor							Observed result
	A	B	C	D	E	F	G	
1	0	0	0	0	0	0	0	s
2	0	0	1	0	1	1	1	t
3	0	1	0	1	0	1	1	u
4	0	1	1	1	1	0	0	v
5	1	0	0	1	1	0	1	w
6	1	0	1	1	0	1	0	x
7	1	1	0	0	1	1	0	y
8	1	1	1	0	0	0	1	z

The above eight combinations of choices provide for high efficiency in that differences between the two choices for each factor are evaluated by contrasting four of the results against the remaining four. Futhermore, the average of all aight results introduces each choice just four times into the average and gives equal weight to all the alternatives.

The importance of giving equal weight is easily recognized by the experimenter in simple situations but apparently overlooked in more complicated settings. The experimenter may have investigated the effect of substituting one calibrated rod, R_0, for another rod, R_1, no other factor being explored. Sufficient repetitions with each rod discloses an unmistakable small difference between the averages for each rod. At this point the investigator, lacking any reason to favor one rod over the other, has no hesitation in giving equal weight to the two sets of results and takes as his *best value* the mean of the two sets.

Suppose the above work has been conducted at temperature T_0. The two series of measurements may be identified by the labels R_0T_0 and R_1T_0. Let the investigator now undertake a third series of measurements using rod R_0 at another temperature T_1. Denote this series by R_0T_1. Clearly the temperature effect (if any remains after proper allowance) is estimated by taking the differences between the averages for R_0T_0 and R_0T_1. Suppose that after due allowance has been made in the computation, the small difference between the results at the two temperatures in also greater than would be expected considering the precision of the work. The investigator should give equal weight to the results at the two temperatures just as he would to the results with each rod. If the mean is taken of the three averages associated with R_0T_0, R_1T_0 and R_0T_1, clearly twice as much weight is given rod R_0 as rod R_1 and twice as much weight to the results at temperature T_0 as to the result at temperature T_1. Indeed, in this awkward combination of choices, equal weight can only be achieved by discarding the average for R_0T_0 and taking the mean of the remaining pair of averages.

Few experimenters are aware that if they were willing to run just one more series of measurements they could not only double the data back of each comparison but also include a third factor on the same terms. Perhaps, in addition to the rod and temperature changes, the work might be extended to examining the effect of changing the voltage from V_0 to another voltage, V_1. The combinations :

$$\begin{array}{ccc} R_0 & T_0 & V_0 \\ R_0 & T_1 & V_1 \\ R_1 & T_0 & V_1 \\ R_1 & T_1 & V_0 \end{array}$$

visibly permit the evaluation of the effect of changing rods by contrasting the last two with the first two. In each pair both temperatures and both voltages have been used so contributions from these factors cancel out. Similar considerations apply to the evaluation of the temperature and voltage effects. Finally the mean of the four series gives equal weight to all alternatives. Unless this is done, another worker using the identical choices in some other combinations cannot expect to converge upon the value reported by the first worker.

The design just given for three factors is particulary neat. Suppose instead of three there are four factors. The statistician will immediately recall the Yates design with seven factors and propose that three of the seven factors be treated as dummy factors. There is, however, the disadvantage of requiring eight series of measurements ; three more than necessary to provide for the unique evaluation of the effects of changing four factors. This enlargement of the experimental program should be avoided, if possible, because the alteration of the equipment may involve considerable time and effort. A change in the apparatus may require much care, as in levelling or making sure there are no leaks in a vacuum system. The restriction of all the factors to just two choices will also be an undesirable limitation. As matters stand, statisticians do not have a collection of designs to meet these requirements.

EXAMPLE WITH THREE EXPERIMENTAL FACTORS

Imagine that an investigator can easily provide three choices, 0, 1 and 2, for each of two factors, and two choices, 0 and 1, for a third factor. The zero choice for each factor is taken as a reference set. An experimenter will usually investigate each factor in turn by conducting the six trials shown in Table 2.

First factor A is explored holding all other factors constant. Then factor B is tried at two new levels and finally the second choice for C is investigated. The effect of changing a factor is looked for by comparing the appropriate average with the average obtained for the reference set $A_0B_0C_0$. Note that the difference between the two averages will have $\sqrt{2}$

times the error associated with a single average. The ability to detect small effects will increase with the number of measurements on which each average is based. If there are as many as 15 measurements for each average, the investigator can be reasonably sure of detecting an effect that is as large as the standard deviation of a single measurement. A large number of degrees of freedom ($14 \times 6 = 84$) are available for estimating this standard deviation.

Table 2. Conventional three factor program.

Trial No.	Factor			Observed average
	A	B	C	
1	0	0	0	u
2	1	0	0	v
3	2	0	0	w
4	0	1	0	x
5	0	2	0	y
6	0	0	1	z

There are two comments to be made regarding the program outlined in Table 2. First a different selection of the six combinations would give the same chance of detecting an effect using ten in place of 15 repeat measurements for each combination. This holds for the three choice factors. There is an additional gain for the C factor. Or, alternatively, if the 15 measurements are retained, still smaller effects will become detectable. The second comment concerns the "best value" or concensus for the final result. Obviously if the mean of the six averages is taken the initial conditions of the reference will be very heavily weighted. In order to give every choice an equal voice in the final result a weighted mean of the six averages must be secured. The proper weighted mean, to give all choices equal representation, is to take one sixth of :

$$-5u + 2v + 2w + 2x + 2y + 3z.$$

The above weights introduce the different choices for the factors into the weighted mean in the manner shown in Table 3.

The tabulation in Table 3 shows that when the six averages, u, v, w, x, y and z, are weighted as shown, the final result gives equal weight to the three choices for the A factor, the three choices for the B factor and the two choices for the C factor. Unfortunately much of the advantage of the repeat measurements is lost. This particular weighted mean has a precision error $\sqrt{1.39}$ times as large as the average for a single combination ! A different program would give results in which the weighting factor for each average is unity. The unweighted mean of the

six averages has a precision error of $\sqrt{1/6}$ as large as the average for a single combination. This better selection of combinations gives nearly a three fold improvement. Of course the experimenter can simply average his six "change one thing at a time" results and obtain a precise estimate of a biased result and this does happen.

Table 3. Weighting factors to obtain unbiased estimate for program in Table 2.

	A			B			C	
Weight	0	1	2	0	1	2	0	1
-5u	-5	0	0	-5	0	0	-5	0
2v	0	2	0	2	0	0	2	0
2w	0	0	2	2	0	0	2	0
2x	2	0	0	0	2	0	2	0
2y	2	0	0	0	0	2	2	0
3z	3	0	0	3	0	0	0	3
Total	2	2	2	2	2	2	3	3

The selection of six combinations that is more sensitive to detecting effects is shown in Table 4.

Table 4. Statistical design for three factor program.

Trial No.	Factor			Observed average
	A	B	C	
1	0	0	0	u
2	0	1	1	v
3	1	0	1	w
4	1	2	0	x
5	2	1	0	y
6	2	2	1	z

Inspection shows that an unweighted mean gives equal weight to all the factor choices. The two diagrams in Figure 1 show the conventional "change one thing at a time" selection in Panel 1 and the statistical design in Panel 2. If there is hesitancy on the part of the experimenter in changing two factors at once, the hesitancy may be partially overcome by observing how much better the design in Panel 2 samples the expe-

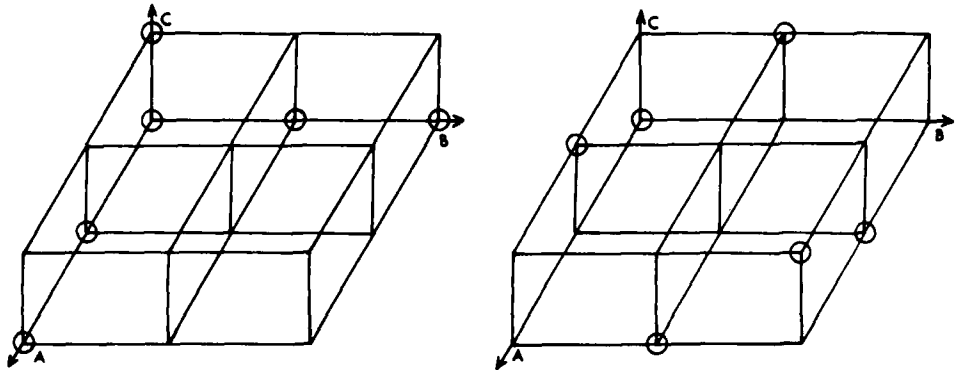


Figure 1 - The circles show selections of six combinations from the 18 combinations provided by three choices for both A and B and two choices for C. Panel 1 shows the conventional selection ; Panel 2 an alternative selection.

rimental region. Those who question whether the effect of changing from A_0 to A_1 is virtually independent of the choices for B and C should study Panel 1 carefully and answer the question as to what sort of information the conventional program would give if the effect of changing A did depend on the choices for B and C.

There remains the task of estimating the effects of the various choices using the statistical design. The estimate of the effects of the various choices for A, B, and C involves a weighted mean of the six observed results. The sums of the results, weighted as indicated in Table 5, should be divided by three in every case. The variance of the comparisons, for the three choice factors, is two thirds that of the "change one factor at a time" procedure, and for the two choice factors the variance is reduced to one third.

Table 5. Estimating effects for design given in Table 4

Observed result	Weighting factors for comparisons				
	A_0-A_1	A_0-A_2	B_0-B_1	B_0-B_2	C_0-C_1
u	2	1	2	1	1
v	1	2	-2	-1	-1
w	-2	-1	1	2	-1
x	-1	1	-1	-2	1
y	-1	-2	-1	1	1
z	1	-1	1	-1	-1

SOME EXAMPLES OF EXPERIMENT DESIGNS

In this section six proposed designs are listed that may find immediate use. Under ordinary circumstances not more than ten different combinations will be studied. The number of choices for a factor will generally be two or three and rarely four. The total number of designs needed is consequently rather small. In any particular case there exists a large number of selections of subsets from the complete factorial. It is not always easy to determine whether the best possible choice among these subsests has been made. The selections listed here may not be the best but they do represent a marked improvement over the usual programs. The addition of one or two extra combinations will, in some cases, provide a much better design with a marked improvement in efficiency.

In the following tabulation the factor choices are denoted by 0, 1, or 2. The usual convention for indicating the number of factors and choices is used. Thus $2^4 3$ indicates that there are four factors each with two choices and one factor with three choices. Two designs are listed for $2^4 3$, one of them with an extra combination.

2^4 in 5 combinations

A	B	C	D
0	0	0	0
0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	0

2^5 in 6 combinations

A	B	C	D	E
0	0	0	0	0
0	1	1	1	0
1	1	0	0	0
1	0	1	1	0
0	0	1	0	1
1	0	0	1	1

$2^3 3$ in 6 combinations

A	B	C	D
0	0	0	0
1	1	0	1
0	1	0	2
1	0	1	2
0	0	1	1
1	1	1	0

$2^4 3$ in 7 combinations

A	B	C	D	E
0	0	0	0	0
1	1	1	1	0
0	1	0	1	1
1	0	1	0	1
0	0	1	1	1
0	1	1	0	2
1	0	0	1	2

$2^4 3$ in 8 combinations

A	B	C	D	E
0	0	0	0	0
1	1	1	1	0
1	1	0	0	1
0	1	1	0	1
1	0	0	1	1
0	0	1	1	1
0	1	0	1	2
1	0	1	0	2

$2^4 3^2$ in 9 combinations

A	B	C	D	E	F
0	0	0	0	0	0
1	1	0	0	1	0
0	0	1	1	2	0
0	1	0	1	0	1
0	0	1	0	1	1
1	0	0	0	2	1
1	0	1	0	0	2
0	0	0	1	1	2
0	1	0	0	2	2

DISCUSSION OF PHYSICAL MEASUREMENTS

It would not be altogether surprising should two laboratories determine a physical constant by two entirely different methods based on different principles, if the results obtained showed a disagreement considerably beyond that anticipated from the internal precision within each laboratory. Either or both of the *methods* may have a systematic error through some defect in theory. It is more surprising to find substantial disagreement when two laboratories use the same procedure and use equipment that differs only in minor ways, for example, dimensions. The only plausible explanation appears to lie in the uncertainties in the calibration of various component parts of the apparatus and in the instruments used to record the relevant operating conditions. Inevitably the investigator has to depend on other workers to provide these indispensable calibrations and to accept, along with the calibration, some statement regarding the accuracy of the calibration. If only one of each of the component parts is available and only one specimen of each necessary instrument at hand, the investigator has no check on the claimed accuracies. Even if a choice does exist, the detection of discrepancies requires a considerable number of measurements with each choice. The individual discrepancies must be kept small because the final result reports the net *sum* of the systematic errors associated with the various components.

If only single choices are available, the experimenter can do no more than estimate the error in the final result from the information available to him regarding the uncertainties in the individual components. If there are two or more choices, the consistency of the results with the two choices furnishes a check on the claimed accuracies. Furthermore, the most troublesome components will be established on the basis of experimental evidence. If the state of the art stands in the way of any immediate improvement of a particular troublesome component, at least

the opportunity exists to procure several choices for such a component and obtain the benefit of an average.

The most important return to the experimenter from the use of a number of different combinations lies in the more realistic estimate of the error in the final result. The $2^4 3^2$ design allows six components to be studied, four with two choices and two with three choices. There are altogether 144 possible experimental combinations given these choices. If no choices had been available, the investigator would base his report on some one of these 144. With each of these 144 combinations there is associated a net systematic error which is the algebraic sum of the systematic errors in the six choices actually employed. There is an 0.5 chance that two laboratories will differ even in the signs of these sums. And, of course, sums of n errors have a wider dispersion than the individual errors. If the experimenter elects to try nine of the 144 possible combinations, he has an opportunity to see for himself the discrepancy that could happen when he compares his result with the result from another laboratory. In another laboratory, perforce, different choices for all the components will be used. A realistic estimate of the uncertainty in the final value can be obtained from the dispersion exhibited by the nine results associated with the nine combinations. (It is assumed here that no one component has an uncertainty that dominates all others).

This design makes it possible to detect considerably smaller differences between components for the same number of measurements. Most important, this particular selection of nine from the 144 possible combinations will provide a final average of high efficiency that gives equal weight to all the choices available for the six components. The net errors associated with the nine results also undergo an averaging out in the mean so that the systematic error in the final result should be substantially reduced. Certainly this program is no less novel to the statistician than to the experimenter because the appropriate estimate of error is, in fact, based on the mean squares associated with the main effects of the several factors.

LITERATURE CITED

[1] YATES F. - "Complex experiments". Jour. Royal Statistical Society, Supplement. 2 : 181-233, 1935. (See page 210).

DISCUSSION

J. NEYMAN : As Dr. Youden has indicated, there are still a great many domains of scientific research in which the statistical principles of experimentation are not yet generally accepted.

Astronomy, meteorology and, partly, medicine are good examples.

In these circumstances, a publication of a collection of examples of studies in which unreasonable results were obtained because of the neglect of some detail might be useful.

M. W. J. YOUDEN : In reply to Professor Deming.

Dr. Youden admitted that the progress of science sometimes reveals that a "constant" is not a constant.

Example : some atomic weights depended on the geographical source of the element. The old "constant" was replaced by the atomic weights of the elements.

In reply to Professor Mahal.

Agreed that laboratory "effects" were important and mentioned that national laboratories now cooperated by intercomparing the same objects first in one country and then in another country.

Mr. FINNEY : I want to mention the Plackett-Burman designs for estimating main effects when all interactions are zero. By contrast, Dr. Youden wants to estimate the general mean, when interactions are zero and the main effects are not of intrinsic interest.

The design in Table 4 is of course a simple 1/3 replicate of $3^2 \times 2$, or one block of Yates's confounding schem. On page 11, the designs proposed are not all perfectly balanced over levels, so that the simple mean would be biased relative to permutations of symbols for levels. Is Dr. Youden aiming simply at minimum bias for a limited number of assemblies, or is he prepared to demand unequal weighting of means ?

M. BOSE : Designs for determining main effects and means in the case when there are no interactions, and when different factors are at different levels are being worked out at the Research Triangle Institute, North Carolina, U.S.A.

Reprinted from: Colloques Internationaux du Centre National de la Recherche Scientifique No. 110, le Plan d'Experiences, 1961, pp. 115-128.

3. Interlaboratory Tests

Foreword

In conducting an interlaboratory test, we usually have one of three purposes in mind:

A. troubleshooting, or audit of the comparability of measurements,
B. evaluation of a test method, or
C. extension of a measurement process from a primary laboratory to other standards laboratories.

According to the purpose to be emphasized in a particular round of tests, the approaches to the problem are necessarily different.

Youden's several papers on graphical analysis (3.1, 3.2) and his ranking scores procedures (3.3, 3.4, 3.6) are designed to locate and identify sources of trouble through graphical representation which is easily interpretable. In addition, he suggested that a test procedure must be checked out for "ruggedness" to disclose factors that may change from laboratory to laboratory. Indeed the procedure described in paper (3.4) has come to be called "Youden's ruggedness test." Youden's main emphasis is on troubleshooting through experimental design. The applications of his method are extremely effective once the procedures and methods of measurement are well defined.

Mandel and Lashof (3.7) approached the problem from a somewhat different point of view. Given an established test method, they aim to interpret the results through a "linear" model. The analysis segregates the total variability into three components: one due to replication, one due to scale (e.g., calibration of instruments), and one due to variability between laboratories. The emphasis is on the evaluation of a test method and on the quantitative estimation of the effect of these components.

But why should there be any between-laboratory differences? Why can't we eliminate this source of variability altogether? This question must be answered before we can use the Cameron-Pontius philosophy for mass measurement (1.1, 7.1) to demonstrate that accuracy levels attained at NBS can be realized by other primary laboratories throughout the nation. This goal is still far away. Current studies deal with mass and volt calibration and the results are encouraging. As time goes on, procedures will be developed for assuring that a measurement process is not only independent of time and conditions at a single laboratory, but also independent of location. The design and analysis of interlaboratory test procedures will play an important role in providing standards of constancy and compatibility.

(Reprinted from INDUSTRIAL QUALITY CONTROL, Vol. XV, No. 11, May 1959)

Graphical Diagnosis of Interlaboratory Test Results

W. J. YOUDEN

National Bureau of Standards, Washington, D. C.

Introduction

Interlaboratory or round robin programs to evaluate the performance of test procedures will always be with us. New materials require new tests. New, and hopefully better, test procedures are developed for old products. Test procedures are used to ascertain whether a product meets the specification set down for the product. A double problem confronts the producer. There is bound to be a certain amount of variation in his product. And there is bound to be variation in the test results made on a given sample of the product. The impact of the errors of measurement associated with the test procedure is obvious because half the tests made on a product that just meets specification will rate the product below specification.

Test Procedures and Production Costs

It is customary to manufacture purposely a product that exceeds specification in order to allow for testing errors. The larger these testing errors, the greater the excess quality that must be built into the product to insure the acceptance of nearly all lots that are in fact equal to or better than the specification. The manufacturer already has to contend with variation in the process. Considerable saving in manufacturing costs can be affected by reducing the margin between the quality level set for production and that called for in the specification. The savings attainable with improved test procedures are a strong inducement for the improvement of test procedures. Interlaboratory test programs of varying degrees of thoroughness are frequently used to establish the performance of existing procedures.

Missed Opportunities in Interlaboratory Test Programs

Strangely enough modern statistical tests such as the analysis of multifactor studies and the isolation of components of variance have not made the contribution expected of them. Part of this no doubt comes about because these more sophisticated statistical techniques are not too well understood by some of those in the laboratories that run the tests. It is all very well for someone with statistical skill to set up an intricate interlaboratory test program and analyse the data but this still leaves the problem of interpreting the statistical jargon to those directly concerned. Even when this interpretation is undertaken the report is apt to read somewhat along these lines. "Duplicates run by the same operator in the

same laboratory show excellent agreement. Agreement between different operators in the same laboratories is not quite so good, and very poor between results from different laboratories. Results on different days do not agree as well as those obtained on the same day." This is a brief summary of the interpretation that is made after the statistical analysis shows that practically all the F-tests are significant. Unhappily almost all concerned were already aware of the state of affairs just described and want to know what can be done to improve matters. It is just here that statisticians have not risen to the opportunities presented by interlaboratory test programs.

When all is said and done, what we want is rather simple. We want to know whether the test procedure as set forth is capable of yielding acceptable agreement among results from different laboratories. If the results are not acceptable, we would like some specific indication of what is wrong with the procedure. If the procedure appears to be reasonably good but there are some disturbing discrepancies, we would like to know which laboratories are having trouble and if possible why they are having trouble. And most important we should be able to get this information back to the laboratories concerned in such a form that the diagnosis is believed. For only so will these laboratories take any action to correct the difficulties.

Graphical Representation of Results

The graphical procedure is based upon a very simple interlaboratory program. Samples of two different materials, A and B, are sent to a number of laboratories which are asked to make one test on each material. The two materials should be similar and be reasonably close in the magnitude of the property evaluated. This will avoid complications that may arise from differential behavior of the two test materials. A second pair of samples are circulated at a later time if there are only a few participating laboratories. The pairs of results that are reported by the laboratories are used to prepare a graph.

The graph is prepared by drawing the customary x-axis at the bottom of the paper and laying off on this axis a scale that covers the range of results for material A. At the left the y-axis is provided with a scale in the same units that includes the range of results reported for material B. The pair of results reported by a laboratory are then used to plot a point. There will be as many points as there are reporting laboratories. After the points are plotted a horizontal median line is drawn parallel to the x-axis so that there are as many points above the line as there are below it. A second median line is drawn parallel to the y-axis and so placed that there are as many points on the left as there are on the right of this line. Figure 1 shows the seven-day tensile strengths reported by 25 laboratories on two cement samples. Two of the laboratories are so patently separated from the other 23 that they are not used in determining the position of the median lines.

Diagnosis of the Configuration of Points

The two median lines divide the graph paper into four quadrants. In the ideal situation where *only* random errors of precision operate the points are expected to be equally numerous in all quadrants. This follows because plus and minus errors should be equally likely. In any existing test procedure that has come to my attention the points tend to concentrate in the upper right and lower left quadrants. This means that laboratories tend to get high results on both materials or low results on both materials. Here is evidence of individual laboratory

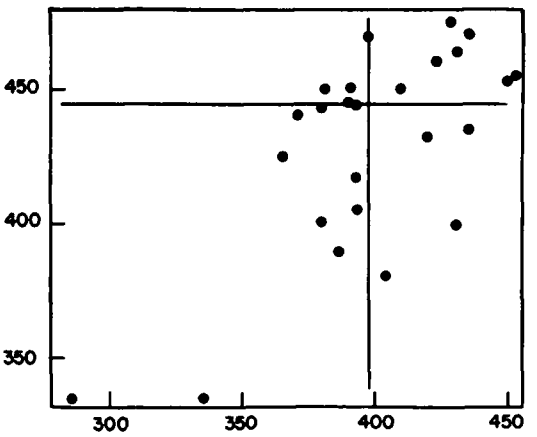


Figure 1—Tensile Strength

biases. There is evidence of this state of affairs in Fig. 1. The more pronounced this tendency to individual bias the greater the departure from the expected circular distribution of points about the intersection of the median lines.

Figure 2 shows 15 points plotted from phthalic anhydride determinations on two paint samples. The points tend to scatter more or less closely along a line approximately bisecting the upper right and lower left quadrants. There is reason to expect the line to make a 45 degree angle with the axes when the same scale is used for both axes and the two materials are sufficiently similar so that the dispersion of the results is about the same for each material.

A test procedure that yields results like those in Fig. 2 is probably in need of more careful description. In its present form the procedure apparently is open to individual modifications that do have an effect upon the results. The procedure rather than the laboratories should be considered as a possible source of the difficulty even though the difficulty is exhibited by a large scatter among the results from the different laboratories. When the points lie closely along the 45 degree line the conclusion may be drawn that many of the laboratories

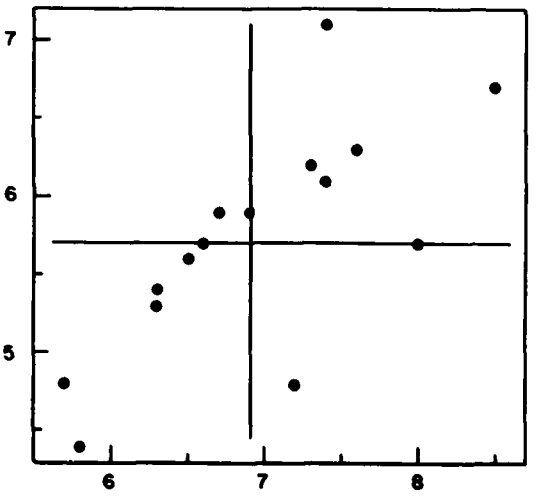


Figure 2—Percent Phthalic Anhydride

are following rather carefully their own versions of the test procedure.

Checking on Sample Variation

There is no possibility of the distribution of points in Fig. 2 arising from lack of uniformity among the samples distributed from each material. If the stock is heterogeneous, some samples will be high, some low, and this will be true for both materials. The pairs of samples distributed to the laboratories will be of four kinds:

high in A, high in B
high in A, low in B
low in A, high in B
low in A, low in B

The four possible combinations have the same probability of occurrence and would result in the test results being nearly equally divided among the four quadrants. Concentration of the points in two quadrants rules out questions of sampling heterogeneity.

On the other hand if there is a roughly circular distribution of points but with a disappointingly widespread scatter, the diagram does not reveal whether this arises from sampling difficulties or poor precision of the test results. If sampling is considered a possible source of difficulty the following modification in the assignment of samples should be tried. If there are 2N laboratories, prepare N *double-size* samples for each material. Carefully mix and divide each double-size sample into two usual size samples.

Double size sample	Laboratory	Samples A	Samples B
1	1	1A	1B
	2	1A′	1B′
2	3	2A	2B
	4	2A′	2B′
N	2N−1	NA	NB
	2N	NA′	NB′

The samples are assigned to laboratories as shown above. It should be possible to mix and divide each double-size sample into two closely matching regular samples. These samples are assigned to a pair of laboratories. If there are sampling difficulties the plotted points should tend to occur in doublets. Two laboratories getting the two carefully mixed halves should check each other and have their points close together. This involves a little extra work in getting out the samples and no extra work for the participating laboratories. If the points corresponding to the two halves of a double-size sample are separated as much, on the average, as points from different double samples, the dispersion cannot be ascribed to sampling. In addition to noting the spacial distribution the projections of the points on the axes may also be used to see whether just one of the materials was heterogeneous.

Interpretation of Out-of-Line Results

So far the large aspects of the diagram have been examined. The individual points can now be considered and in particular those points most distant from the intersection of the median lines. Almost always one or more points are so far out of the picture that it is better not to compress the scale in order to show them. Such points should be ignored in locating the median lines. (See Fig. 1.) The more distant points tend to fall into one or the other of two categories. Either the point is far out and remote from both axes or far out and fairly close to one or the other axis. In the latter case, the result is fairly good on one material and very bad on the other. Examples of such points are found in Figures 2 and 3. Often the explanation is simple—a mistake in typing, or calculation, or some simple blunder that sometimes can be corrected by going back to the records. If the same laboratory shows up in such a manner on succeeding pairs of materials, this implies carelessness on the part of the laboratory. The laboratory can do good work but often does not. Occasionally a laboratory has difficulty with one material and not with the other but this is not likely to occur with similar materials.

Points in the upper right or lower left quadrants that are far removed from the intersection of the median lines and that are not near either axis reflect a tendency to get either high results on both materials or low results on both materials. There are examples in all the figures. The more consistent a laboratory is in its work the more likely its point will lie in the proximity of the 45 degree line. A point far out along this line suggests the possibility that the laboratory concerned has introduced some modification into the test procedure. A laboratory finding itself in this situation should check carefully the prescribed procedure for performing the test and endeavor to locate the cause of the large bias.

All of the above interpretation can be made while keeping anonymous the identity of the plotted points. When circulating a report of the interlaboratory test it might be helpful to circle in red the point belonging to the laboratory in the copy going to that laboratory. That would save the laboratory from consulting its files to locate itself and would display prominently just where the laboratory stood in reference to the whole group. This vivid picturing of a laboratory's position should stimulate the laboratory to some self examination that could hardly avoid having beneficial results.

Estimating the Precision of the Test Procedure

The above discussion does not exhaust the information to be gleaned from this graphical representation. Assuming that the two materials are similar in type and nearly equal in magnitude for the property the dispersion among the results reported for A should be about the same as the dispersion of the B results. In that event the 45 degree line through the intersection of the medians makes possible an estimate of the *precision* of the data. Often an interlaboratory test undertakes to differentiate among the laboratories in respect to precision. Not only does this require large numbers of measurements from each laboratory but differences in precision usually turn out to be unimportant in comparison with bias errors and careless errors. No violence at this stage seems to be done by assuming about the same precision for all the laboratories.

The perpendicular distance from each point to the 45 degree line can be used to form an estimate of the precision. The estimate of the standard deviation of a *single* result is obtained by multiplying the average length of the perpendiculars by $\sqrt{\pi/2}$ or 1.2533. These perpendiculars need not be measured on the graph paper. Instead, write d wn for each laboratory the difference (A−B) keeping track of the signs. Call these differences $d_1, d_2, \ldots d_n$. Calculate $\bar{d}$, the algebraic average difference. Subtract $\bar{d}$ from each difference and obtain a set of corrected differences $d_1', d_2', \ldots d_n'$. The average of the absolute values of these differences when multiplied by $\sqrt{\pi/2}$ or 0.886 gives an estimate of the standard deviation.

TABLE I—Data and Calculations on Percent Insoluble Residue in Cement Reported by 29 Laboratories

Laboratory	Percent Residue A	Percent Residue B	A — B	(A—B) — 0.095
1	0.31	0.22	0.09	—0.005
2	0.08	0.12	—0.04	—0.135
3	0.24	0.14	0.10	0.005
4	0.14	0.07	0.07	—0.025
5	0.52	0.37		
6	0.38	0.19	0.19	0.095
7	0.22	0.14	0.08	—0.015
8	0.46	0.23		
9	0.26	0.05	0.21	0.115
10	0.28	0.14	0.14	0.045
11	0.10	0.18	—0.08	—0.175
12	0.20	0.09	0.11	0.015
13	0.26	0.10	0.16	0.065
14	0.28	0.14	0.14	0.045
15	0.25	0.13	0.12	0.025
16	0.25	0.11	0.14	0.045
17	0.26	0.17	0.09	—0.005
18	0.26	0.18	0.08	—0.015
19	0.12	0.05	0.07	—0.025
20	0.29	0.14	0.15	0.055
21	0.22	0.11	0.11	0.015
22	0.13	0.10	0.03	—0.065
23	0.56	0.42		
24	0.30	0.30	0.00	—0.095
25	0.24	0.06	0.18	0.085
26	0.25	0.35		
27	0.24	0.09	0.15	0.055
28	0.28	0.23	0.05	—0.045
29	0.14	0.10	0.04	—0.055
Average	0.229	0.134	0.095	0.053

The data on percent insoluble residues reported by 29 laboratories are given in Table I and plotted in Fig. 3. There are three points far out along the 45 degree line and one far out on the y-axis. These laboratories were excluded from the calculations shown in Table I. The last column shows the differences between the two results diminished by the difference between the two sample averages. The average, 0.053, shown at the bottom of this column is the average of the *absolute* values, i.e., ignoring the signs. Multiplying 0.053 by 0.886 gives 0.047 as the estimate for the standard deviation of a single result. Probably this is inflated by leaving in the two laboratories turning in the very low results for sample A.

This estimate of the standard deviation for precision leads to the construction of circles (centered on the intersection of the median lines) within which any given percentage of the points can be expected to fall should the laboratories be able to eliminate all bias or constant errors. The multiples of the standard deviation that include various percents of the points are given in Table II.

Thus a circle whose radius is about 2.5 to 3.0 times the standard deviation gives a fair idea of the smallest circle that could be expected to contain nearly all points after the elimination of the constant errors that are causing the points to congregate in the upper left and lower right quadrants. Generally a fair number of points will lie outside such a circle. The laboratories responsible for these points almost certainly have somehow got substantial systematic errors incorporated in their techniques. Multiplying the standard deviation obtained above by 2.45 gives the radius of the circle that should include 95 percent of the laboratories if individual constant errors could be eliminated. This circle is drawn in Fig. 3. Seven further laboratories are outside the circle including the two who got the benefit of the doubt and were retained in the computation. This examination has directed attention to at least six of the laboratories that might well go over their method of making this determination of insoluble residue.

TABLE II—Probability Table for Circular Normal Distribution

Percent of the Points Within Circle	Multiple b of the Standard Deviation
10	0.459
20	0.668
25	0.759
30	0.845
40	1.011
50	1.177
60	1.350
70	1.552
75	1.665
80	1.794
90	2.146
95	2.448
99	3.035

Note: Percent = $100[1 - \exp(-b^2/2)]$

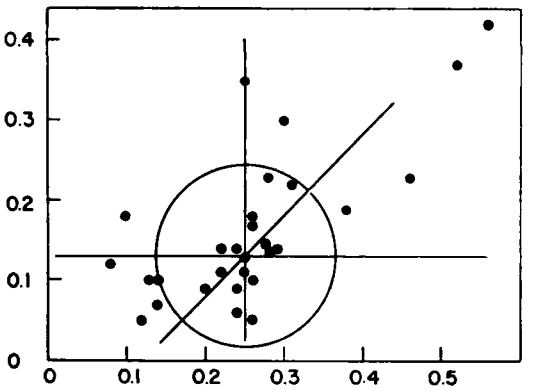


Figure 3—Percent of Insoluble Residue

If the number of laboratories in the program is rather small, the way to accumulate more points is to send the laboratories additional pairs of samples from different materials. A chart is prepared for each pair of materials and the median lines drawn in. The charts are now superimposed so that the points of intersection of the median lines coincide and, of course, the median lines also. All points are then transferred to one sheet of paper with one pair of median lines. As there are only a few laboratories each can be assigned an identifying symbol.

Figure 4 shows the reports made by eight laboratories determining CaO in cement. The laboratories are

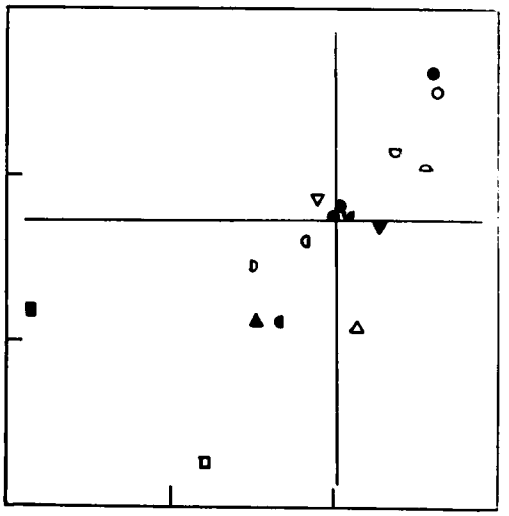


Figure 4—CaO in Cement (interval equals one percent)

identified by symbols. The hollow symbols show the results for the *first pair* of samples. The corresponding solid symbols show the work of these same laboratories on a *second pair* of samples. Few as these data are they serve to indicate the things that we want to know about the test procedure and about the laboratories. Clearly this procedure is one that is vulnerable to individual bias. Two of the eight laboratories appear in the same region for both pairs. The circle laboratory is very consistent—and gets the highest results. The square laboratory gets very low results and is not very precise as shown by the fact that the two squares are separated by a much greater distance than any of the other seven pairs. Using this chart some possibly helpful suggestions could be passed along.

Discussion

The two materials used in this double-sample program were specified to be similar in type and in the magnitude of the property measured. Sometimes the measurement errors are proportional to the magnitude under measurement and this will show up in a greater scatter of the points along one of the axes. Particular types of samples may give trouble in just some of the laboratories. The thorough study of a test method must include consideration of these possible complications. Naturally a more comprehensive interlaboratory test program will be required to explore these aspects of the test procedure. A thorough study in one laboratory usually reveals these complications.

Summary of Advantages of Graphical Diagnosis

The double-sample, graphic analysis scheme described in this article offers a number of advantages.

(1) An unusually light burden is imposed on each laboratory

(2) The graphical procedure greatly facilitates presentation of the results in a convincing manner

(3) No statistical background is required to follow the reasoning and no computations are required to demonstrate the general presence of constant errors and the gross deviations of individual laboratories

(4) A minimum of computation is imposed upon the individual collating the results

(5) The use of a circle of 2.5 or 3.0 σ radius shows the individual laboratories whether or not their method of carrying out the test has in some way become saddled with a substantial constant error

(6) Most important the direction for improvement is clearly indicated

a. A long, narrow ellipse directs attention to a more careful description of the procedure or even to the need for modification

b. Wild points far out near either axis indicate erratic work

c. Wild points far out along the 45 degree line are strong evidence of substantial deviations from the specified procedure

d. General prevalence of constant errors is indicated by a substantial proportion of the points lying outside the 2.5 σ circle

Experience has already indicated that a certain few laboratories are found too frequently in the most distant positions from the intersection of the median. Improved performance from these few laboratories may go far to restore confidence in a test procedure. There is no substitute for careful work in the laboratory.

The Sample, The Procedure, and The Laboratory

W. J. Youden
National Bureau of Standards, Washington 25, D. C.

IN THIS paper the viewpoint is taken that an analytical procedure has an inherent accuracy and precision. True enough, there must be an analyst in a laboratory to put the procedure to work and this implies to some analysts that an inseparable association exists between procedure and operator. A sample is also indispensable, yet there is no hesitation in sometimes attributing the variation in analytical results to a lack of homogeneity in the material furnishing the samples. At other times, often when a reasonable volume of a liquid is sampled, the aliquots used as samples can be considered identical in composition and any differences among the results cannot be charged to the samples.

The role of the analyst, or laboratory, may be revealed when two or more laboratories undertake determinations on samples drawn from the same stock of uniform material. In extreme cases the repeat determinations made by a laboratory cluster closely about the laboratory average without any intermingling of the results from one laboratory with the results from another laboratory. Figure 1 illustrates this point. The open circles represent the results from one laboratory and the solid circles the results reported by a second laboratory. Separation of the results from different laboratories is practically always present to some extent—that is, the separation between results from different laboratories is greater than would be anticipated, considering the agreement among the results obtained within a single laboratory. The reduction, or, if possible, the elimination of these interlaboratory differences is an everyday problem. Here is a major reason why busy analytical chemists turn to statistical techniques for help in resolving the complex of circumstances that surround analytical determinations.

Wrong Operations on Data

Often a study makes available a collection of analytical results obtained under a variety of circumstances. One wrong operation is to take the grand average of all the data and obtain the individual deviations from this average. It matters not whether the simple arithmetic average of these deviations (of course ignoring signs) is reported, or some more sophisticated quantity, such as the standard deviation, is computed. The quantity so reported is almost surely useless, if not downright misleading. Nor will matters be helped if the analyst happens to have available the theoretical or assumed true composition of the material and is able to measure his deviations from the true value. In fact, this usually makes matters worse. I am fully aware that these computations are very generally made, but they are made in the mistaken belief that the simplicity of the calculations ensures a meaningful result.

An illustrative example will clear the ground of erroneous operations on the data. The example is taken from some long ago microanalytical determinations of carbon reported by Power (*1*). Analyst H reported six determinations on pure ephedrine hydrochloride as follows:

59.09, 59.17, 59.27, 59.13, 59.10, 59.14
Av. 59.15

If the deviations are obtained by subtracting from these results the theoretical per cent of carbon, 59.55, the deviations are

−0.46, −0.38, −0.28, −0.42, −0.45, −0.41
Av. −0.40

We are immediately struck by the unvarying minus sign and the relative constancy of these large negative deviations. By accident, in this example, because all the deviations have the same sign, the average of these deviations (−0.40) is informative. It is, in fact, an estimate of the bias or systematic error in the results, and if the sign is retained, we have the *direction* of the bias. The average deviation is not always so kind as to furnish an estimate of the bias. When the signs of the deviations are not all the same,

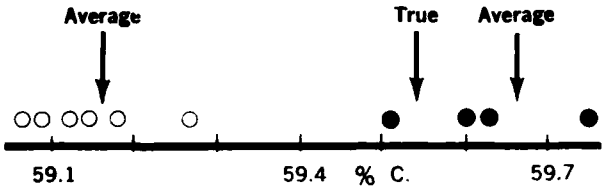
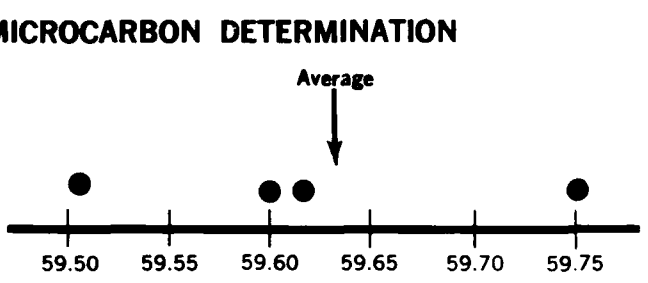


Figure 1

138-23A



Inclusion of extreme values may displace average unduly

the average of the absolute deviations no longer measures the bias—or anything else. Probability statements cannot be made about the above deviations because they all have the same sign. One could state that no matter how many determinations had been made, they would all have given negative deviations from the true composition.

Power listed four of his own determinations that he considered acceptable. His results were 59.51, 59.75, 59.61, and 59.60, with an average of 59.62. Apparently Power avoided whatever circumstances led analyst H to his low results. The ten deviations that would be obtained by taking differences from the average of all ten results tell nothing useful. The deviations reflect a confused mixture of random errors and systematic errors. Even the average used clearly depends upon the relative numbers of determinations provided by the two analysts. If the theoretical composition is used, the deviations visibly consist of two groups with no intermingling. Statistical statements for such heterogeneous deviations are meaningless. It is more informative to state for each analyst the departure of his average from the theoretical composition, for each to give an estimate of his precision using the deviations from his own average.

When the magnitude of the systematic error is comparable to the random errors associated with precision, a predominance of the deviations from the true value will have the same sign. When a random error of opposite sign and somewhat larger than the systematic error comes along, the net result is to give a sign opposite to that shown by the majority. The best evaluation of the random errors exhibited by the above six results is obtained by using the deviations of the individual determinations from the average of all the results. The deviations, −0.06, 0.02, 0.12, −0.02, −0.05, and −0.01, must sum to zero and should show a reasonably equal partition between plus and minus signs.

The *estimate* of the standard deviations associated with the laboratory in which analyst H made his determinations is given by $s = \sqrt{\Sigma(\text{dev})^2/(n-1)}$ or 0.065. The estimate of the bias, −0.40, is about six times as large as s. A random deviation (either plus or minus) of this magnitude is extremely unlikely. Hence all the signs of the deviations are the same. As the ratio of the bias to the standard deviations gets smaller, there is more likelihood of a mixture of signs. Table I shows for various values of this ratio the expected division of the signs of the deviations from the true value.

This particular example was chosen to bring out clearly the two concepts of a systematic component of error and a random component of error. It may be, that in as clear cut a situation as this one, few would go astray. But it must be remembered that there is a continuum extending from very large obvious biases down to very small biases. The values computed from the data should correspond to meaningful chemical quantities. The separation of bias from random errors is indispensable to an efficient approach to the improvement of analytical procedures.

Statisticians have unwittingly contributed to the confusion when they remark that the divisor for the sum of the squared deviations must be one less than the number of measurements, because the deviations are measured from the average rather than the true value. The statistician and the chemist refer to quite different things when they speak of the true value. The chemist has in mind the actual correct composition. The statistician means the value that the average of the results would approach with an indefinite increase in the number of determinations made under the same conditions. In other words, the statistician's true value includes the systematic error, if any.

True Composition Unknown

If the true composition is not known, the estimation of the magnitude of a systematic error in the results is not so easy but in some situations not impossible. If the systematic error in the determinations is the same over a considerable range of sample weight (or volume), the systematic error may be estimated by plotting the actual measured quantity against the sample

weight. The measured quantity may be the weight of a dried precipitate or the milliliters used in the titration. Clearly if one sample weight is twice the weight of another sample, there should be twice as much precipitate or twice as many milliliters of reagent used. If there is a systematic error that is independent of the sample weight, all the results should be high (or low) by the same amount. A straight line fitted to the points will not go through the origin, as it ought to, but will intercept the y-axis. The intercept is an estimate of the systematic error. This device fails if the systematic error is proportional to the amount taken for analysis.

While it may be difficult to estimate the magnitude and sign of the systematic error, the demonstration that systematic errors are present is all too easy. If two laboratories report a number of analyses on the same material, any difference that can be established between the laboratory averages is evidence that one or the other or both sets of results are afflicted with a systematic error. It was shown above that any attempt to describe such joint collections of data by a single statistical unit is bound to be misleading.

The evaluation of analytical data is greatly simplified if it is assumed that the participating laboratories have the same precision. The basis for this assumption is that apparatus, equipment, and analyst training are highly standardized and of high quality. Weighings, titrations, instrument readings, and the like are likely to be made with about the same reproducibility. Usually if there are differences in apparatus or technique, these concern matters that do not contribute appreciably to the precision. Weighing errors, for example, are usually a minor consideration, so that little consequence comes from one laboratory using a balance with twice the sensitivity of the balance used in the other laboratory. Thoughtful consideration of the steps in an analytical procedure soon leads to the conclusion that differences between laboratories in regard to equipment, reagents, or in procedures are more likely to lead to systematic errors than to changes in precision.

The most obvious source of a systematic error is a deliberate or unwitting departure from the prescribed manner of carrying out the procedure. Chemists are individuals; they have their favorite precautions, short cuts, and prejudices.

Table I. Division of Plus and Minus Signs of Deviations from True Value Depends on Ratio of Systematic Error to Statistical Deviation

Systematic Error / Standard Deviation	Division of signs of Deviations, %	
2.0	97.7	2.3
1.5	93.3	6.7
1.2	88.5	11.5
1.0	84.1	15.9
0.8	78.8	21.2
0.6	72.6	27.4
0.4	65.5	34.5
0.2	57.9	42.1
0.0	50.0	50.0

If a chemist faithfully follows his own routine, his own analyses check each other extremely well. The same will be true for a chemist in another laboratory. His internal checks are no doubt just as good as those obtained in the first laboratory (same precision) but the results, as a group, may reflect the established practice of the laboratory. Similarly reagents in the two laboratories may be from different sources, or lots, or of different ages. All determinations run with a given set of reagents may show excellent internal agreement but average out at a value removed from the aver-

---

W. J. Youden, a statistical consultant at the National Bureau of Standards for the past 12 years, is an unusual combination of analytical chemist, chemical engineer, and statistician.

Although an Australian by birth, he came to the U. S. at an early age. He received his B.S. in chemical engineering from the University of Rochester (1921), and his Ph.D. in analytical chemistry from Columbia University (1924). His thesis concerned a new method for the gravimetric determination of zirconium. In 1937 he held a Rockefeller Fellowship at the University of London.

He joined the staff at the Boyce Thompson Institute for Plant Research in 1924. During the following 24 years, he did research on such topics as tobacco virus, isoelectric points, soil sampling, sugar analysis, seed treatment, pH methods, agricultural field trials, and greenhouse fumigation. As a result of some of those studies he became involved in statistical approaches and in particular to the design of experiments.

He put his statistical skills to work in a completely different area during World War II when he served as an operations analyst in the area of bombing accuracy with the Army Air Forces overseas (1942 to 1945). He also was an operations analyst for the Rand Corporation in 1947.

He joined NBS in 1948 as a statistical consultant. His major interests are the design and interpretation of experiments and the application of statistical techniques in analytical chemistry.

He has served as a visiting professor at the North Carolina State College (1951, 1954, 1955) and as a professor at the University of Chicago (1959). He has also given continuation lectures on the design of experiments for the Philadelphia and New York Sections of the ACS (1951 and 1954, respectively) and has been on seven speaking tours for the ACS and one for the Canadian Institute of Chemistry. He has served as a statistical consultant on several government boards, committees, and councils.

He has been a member of the ACS for 40 years. He is also a member of Sigma Xi, Phi Beta Kappa, and Phi Lambda Upsilon and is active on several ASTM committees. For his work on Youden squares, chain blocks, linked blocks, and partially replicated Latin squares, he has been honored by statisticians who have made him a Fellow of the American Statistical Association, a member of the International Statistics Institute, and a titular member of the Commission of Technology and Expression of Results of the Analytical Section of the International Union of Pure and Applied Chemistry.

age of determinations made with another set of reagents. Pieces of equipment may differ in their zero settings and introduce different biases without in any way altering the precision of the readings. Geographical location sometimes involves fairly persistent humidity differences between laboratories and this may be a reason for the difference between laboratory results.

Finally there is an abundance of evidence that different laboratories have different systematic errors for a given procedure. Little convincing evidence exists of differences in precision. Of course each laboratory likes to believe that it does particularly precise work. Sometimes this belief is bolstered by a too enthusiastic culling of results and running of extra repetitions until a "satisfactory" agreement is obtained. Leaving aside any spurious apparent differences in precision generated in this manner, it seems fair to conclude that laboratories with equivalent equipment and personnel achieve about the same precision.

In any event, it takes a lot of determinations to make a convincing case for differences in precision. Suppose two laboratories each make ten determinations and an estimate is made of the standard deviation for each laboratory. One of the estimates of the standard deviations must be at least twice the other estimate to provide reasonable grounds for the suspicion that there is a real difference in the quality of the work. Suppose that one laboratory does regularly turn out work that has a standard deviation one half as large as that associated with the regular work of another laboratory. If each laboratory submits 20 repeat runs, there is only about a four out of five chance that this actual difference will be reflected convincingly enough in the data to warrant the conclusion that the laboratories differ in precision.

A more vivid illustration of the difficulties in the way of discriminating among laboratories is afforded by the following comments. We assume that six laboratories all have identical precision. The laboratories report five determinations apiece and the standard deviations are calculated. Then we should not be surprised if the ratio of the largest estimate to the smallest estimate of the standard deviation is as much as 5.4. Even if the estimates are based upon ten repeat determinations, the ratio may reach 2.8 purely from the chance distribution of the deviations. If ten, instead of six, laboratories participate, the ratios are 6.7 and 3.1. The nature of measurement is such that, even under the ideal conditions of assumed normality and absence of gross errors, any measure of precision is subject to large sampling variation. Unless there is clear evidence to the contrary, the best procedure is to combine, in the proper way, the several estimates of precision and award this value to all participating.

The combination of the estimates is easily effected by adding together the sums of the squared deviations available from the several sets of results and dividing by the sum of the divisors previously employed. The deviations for each set must be measured from the average of the laboratory (or group) from which the data originate. The six results by analyst H and the four results by Power give the following pooled estimate of the standard deviation:

$$s = \sqrt{\frac{0.0214 + 0.0295}{5 + 3}} = 0.080$$

The remarks about apparent and not real differences in precision also apply to different sets of data accumulated *within* one laboratory. Suppose that there are two sets of measurements, each made up of three repetitions. Perhaps these sets were made on different days. If the range, or spread, for one set is twice that of the other, one cannot conclude on this evidence alone that one set of measurements is more precise than the other or that more confidence may be placed in the average of the set with the smaller range. Assuming that, as far as the analyst knows, there was no change in the circumstances, there is no reason to expect a sudden real change in precision. The analyst should take the view that a given procedure, in competent hands, has an inherent precision which can be ascertained. Individual small sets of data will inevitably give estimates of the standard deviation that show considerable variation. This variation in the individual estimates of the standard deviation is natural, however surprising it may seem. Once sufficient repetitions have been accumulated, say 30 or more pairs of duplicates on samples not too widely spread in content of the element, an estimate of the standard deviation can be obtained that should be used in place of any estimate based on some small set of data. Of course, something can go wrong and sometimes does. There are statistical criteria for suspecting out of line results. If the difference between a pair of duplicates is exceptionally large, this is taken as evidence of a mishap. In that event additional determinations are in order.

Once it is accepted that differences in precision between laboratories can be forgotten because, if present, they are probably minor differences anyway, the way is open for a revealing examination of the data. In any event the evidence is conclusive that differences in the systematic errors are the major source of disagreement among laboratories. Certainly, if this were not the case, the whole edifice of standard samples would be without value. Obviously the use of a standard sample to check out a procedure can in no wise alter the *precision* of the analytical work. A standard sample may direct the attention of the analyst to the need to go over his procedure. Rarely will the measures taken make any difference in the agreement of check determinations. If poor agreement between duplicates were the real trouble, the analyst could use improved agreement between duplicates as a criterion of satisfactory results and dispense with standard samples. This is only saying what every analyst knows: Good agreement between duplicates is a necessary but not a sufficient condition for a good procedure.

Systematic Errors

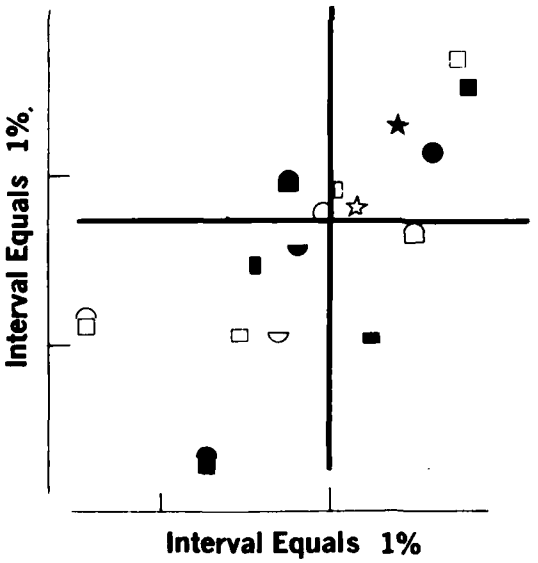
Just as a given analytical procedure may have a certain precision associated with it as a property of the over-all ensemble of operations

involved, so may the procedure itself be thought of as having a built-in systematic error. It is a common remark that this, or that, method tends to give high (or low) results. Obviously gravimetric procedures are vulnerable to low results if the precipitates are too soluble. Very often, in analytical procedures, a blank is specified and clearly this is intended to correct for a systematic error that would otherwise be present. The chemist's goal is to devise procedures that are inherently without any built-in systematic error or bias. It is usually considered sufficient to reduce the systematic error to the point where it is small relative to the precision error.

The systematic error of a procedure is a property of the procedure when performed as specified. Departures from the specified procedure may substantially modify the original bias. Sometimes a laboratory with the best intention of correcting a suspected bias may overshoot and even change the sign of the systematic error. In any event there is no question but that the procedure modifications and the equipment and reagents associated with each laboratory do result in a corresponding gamut of laboratory systematic errors that modify the basic systematic error of the procedure. Considerable advantage follows from accepting this picture of the structure of the systematic error. In the first place the true chemical composition may not be known. All that can be done then is to take as a working reference point the consensus of the participating laboratories. Individual laboratory systematic errors can, in fact, be measured only from this consensus reference point. A particular laboratory that is far out of line may be presumed to have departed from the accepted procedure in a unique way. In the absence of any other guide, the consensus of a reasonable number of laboratories may be taken to characterize the analytical procedure. After all, the laboratories are expected to follow the procedure. At a later date, an opportunity may arise to try the procedure on materials of known composition. Any discrepancy between the true composition and the consensus of the laboratories must be considered a defect in the procedure.

The essential point is that when this way of looking at the systematic error is "simplified" by concentrating attention directly on the difference between each laboratory's own average and the known composition, useful information is lost. Suppose that the systematic error for the procedure is positive and that one laboratory departs from the consensus by a nearly equal negative systematic error. This particular laboratory then has a practically perfect check with the true composition and therefore swears by the procedure. There are some omitted words here. The laboratory swears by the procedure *as carried out by that laboratory*. That does not advance matters at all unless we know, or can find out, in what respects this laboratory departed from the specified procedure. This may be a significant deliberate departure and ascertainable or it may be a chance departure dependent upon the reagents, apparatus, etc., that were used by this laboratory. In all fairness, each laboratory should be judged by its closeness to the consensus, if we have any confidence that the participating laboratories conscientiously tried to follow the procedure in every detail. The discrepancy between the consensus and the true value ought to be charged to the procedure.

The consequence of this point of view is that laboratories close to the consensus deserve pats on their backs. A laboratory whose result departs from the consensus should be called to account even when it *happens* to check the true composition. If the laboratory deliberately departed from the procedure it should share this knowledge, and also simultaneously admit that it did not adhere to the agreement to test the procedure as given. If every laboratory departs capriciously from the procedure as specified, then the whole business of interlaboratory testing might as well be forgotten because no single version of the procedure can be tried



Persistence of systematic errors is shown in two series of analyses run by the same 8 laboratories. Each laboratory is shown by a different symbol. The solid symbols refer to the first series and the open symbols the second series

142-29A

out. If the laboratory has no reasonable explanation to offer for the good check it got, when the consensus of all was clearly not a check, there seems no more reason to congratulate this laboratory than a laboratory that had an equally large deviation from the consensus but in the opposite direction. After all, if chance is operating in the events that introduce laboratory systematic errors, maybe the chances of a plus or negative systematic error are not too different. So one laboratory, judged by the true composition, looks very good, another very bad when perhaps both laboratories have substantial defects in their reagents or apparatus.

When all, or nearly all, the results from a particular laboratory deviate in the same direction from the known composition, the evidence of a systematic error in the results is unmistakable. The advantage of remembering the possible, and likely, composite character of the systematic error, lies in the steps that may be taken to achieve better results. The procedure may require modification. Certain laboratories may need to mend their ways. The desired end is one where all the laboratories cluster closely about their consensus combined with close agreement of the consensus with the known composition. In fact, it can hardly be maintained that an agreed upon procedure exists unless the laboratories can achieve good agreement among themselves around *some* value. Once this stage has been reached, it will improve the chances of successfully locating the cause and remedy for a discrepancy between consensus and true value.

Separation of Systematic and Random Errors

Very few data suffice to demonstrate the presence of individual systematic errors for laboratories and to provide an estimate of their common precision (*2–4*). Two fairly similar materials, not very different in percentage of the element to be determined, will be required. These conditions are stipulated because the precision as well as the systematic error may depend on the per cent of element present and possibly be changed if interfering substances are present. Only one determination is necessary on each material by each of a number of laboratories. If duplicates are run, the averages will be used. Let the materials be designated X and Y. The laboratories are numbered 1 to n, and the results symbolized as $x_1, y_1; x_2, y_2; \ldots; x_n, y_n$. A pair of coordinate axes should be drawn on a piece of graph paper. A scale of values is laid off on the x-axis covering the range from the lowest value reported for X to the largest result. Using exactly the same unit, the scale of values on the y-axis must cover the range from the lowest value for Y to the highest result. Usually the scale is so enlarged that the smallest division on the graph paper corresponds to one unit in the last place of the values reported.

The pair of values furnished by a laboratory determines the location of a point on the graph paper. There will be as many points as there are participating laboratories. A horizontal line is located through the average (consensus) of the values reported for Y and a vertical line drawn through the average of the values reported for X. These two lines divide the graph paper into four quadrants. The pair of deviations from the averages, associated with a laboratory, must be either $++$, $+-$, $-+$, or $--$, and these correspond to the four quadrants just formed. If plus and minus deviations from the average of each material are equally likely, then the four combinations, $++$, $+-$, $-+$, and $--$, are equally probable so that, in theory, equal numbers of points should fall in the four quadrants. This distribution of the points would not be changed even if the laboratories did have different precision, because the signs, and not the magnitudes, of the deviations determine the quadrant getting the point.

Examination of scores of such charts has shown in almost every chart an unequal division of points among the quadrants. Two of the quadrants, the upper right corresponding to $++$, and the lower left, corresponding to $--$, contain a majority of the points. The explanation for such a departure from theory is immediate. If a laboratory does have a systematic error, this error, by definition, appears in both the result for X and the result for Y. While the random errors may be of opposite sign, the deviations will be converted to the same sign if a large enough systematic error is added to, or subtracted from, each random error. The results reported by the laboratories show only the net remaining after random and systematic errors have been combined. The signs give the show away and the surplus of points in the $++$ and $--$ quadrants is graphic testimony of the presence of systematic errors.

Analysts like to dream of a world in which only random errors exist, and small ones at that. Consider the contrary world where perfect precision exists but each laboratory has persistent individual systematic errors. This would mean that if a laboratory's result for X is higher by 0.10% than the consensus for material X, then on material Y it will be exactly 0.10% higher than the consensus for Y—exactly the same amount higher on both materials because of perfect precision (sampling errors assumed not present). In this contrary world all the points would lie precisely on a 45° line passing through the point where the horizontal and vertical lines intersect. Perfect location of all points on such a line has not been observed, but some distressingly near approximations have been encountered.

Most interlaboratory studies yield plots that are intermediate in character between the two extremes of equal numbers of points in the four quadrants and all the points in the $++$ and $--$ quadrants. The points scatter in an approximate ellipse whose long axis is the 45° line through the point corresponding to the averages for X and Y. The larger the systematic errors, relative to the precision error, the more elongated and thinner the ellipse will be. When the points do straggle more or less closely along the 45° line, the evidence for an unsatisfactory procedure is conclusive. Possibly the procedure is in-

adequately described and is so vulnerable to individual interpretation that, as a group, the laboratories are having trouble. On the other hand, if a substantial majority of the points are clustered in a fairly broad ellipse with only a few points far out along the 45° line (either in the ++ or −− quadrants), there is a strong suspicion that the more remote laboratories have their own unique way of making the determinations.

An excuse often advanced by a laboratory with an out of line result is the claim that it got a nonrepresentative sample. This claim is considerably weakened when the laboratory's point is far out and near the line, because now the laboratory has to claim nonrepresentative sample for both materials, and, furthermore, departing in the same way. An even stronger objection can be put forward against this claim. If the materials sampled are not uniform, then, in taking the samples of X, half of the samples will be high and half low. This is also true for material Y. The two samples sent, quite blind, to a laboratory may be high in both (++); high in X, low in Y (+−); low in X, high in Y (−+); or low in both (−−). All combinations are equally likely, so that if the lack of uniformity of the stocks is sufficient to dominate over the systematic errors, then the points should be equally distributed among the quadrants. The argument is now turned in reverse and a lack of equal distribution among the quadrants considered evidence that sample variation is not the problem.

There is a possible ambiguity. Either very poor precision or nonuniform material may lead to an equal distribution among the quadrants. The allocation of samples may be modified to resolve this ambiguity if desired, but the event has not been observed, so means to distinguish between these causes will not be given.

Earlier mention was made that in the event of perfect precision the points would lie exactly on the 45° line. Random errors displace the points from the line. The perpendiculars from each plotted point to the 45° line are a means of estimating the precision of the procedure as revealed by the combined results from the participating laboratories. Designate the lengths of the perpendiculars by $p_1, p_2, \ldots, p_n$. Then an estimate of the common standard deviation is given by

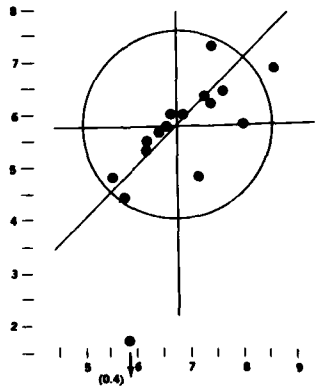
$$s = \sqrt{\Sigma p^2/(n-1)}$$

Some readers may be interested to show that this formula is equivalent to

$$\sqrt{\frac{\Sigma d^2 - n\bar{d}^2}{2(n-1)}}$$

Here each d is the difference between the result reported for X by a laboratory and the result reported for Y by the same laboratory. The algebraic average of these differences gives $\bar{d}$.

Each laboratory provides a perpendicular. Measure the distance along the 45° line from the foot of the perpendicular to the point corresponding to the averages for the two materials. This distance, divided by the $\sqrt{2}$, gives the best estimate of the systematic error of the laboratory measured relative to the consensus of all the laboratories. If the true compositions of the materials are known, they may be used to plot a point. The distance along the 45° line to the true point divided by $\sqrt{2}$ gives an estimate of the systematic error of the procedure as used by the participating laboratories.



The extensive range of systematic errors noted in results by a large number of laboratories all analyzing the same sample of phthalic anhydride indicates the possibility of a faulty procedure

144-34A

Number of Laboratories Required

The small amount of work called for from each laboratory should make it easier to enlarge the number of participating laboratories over the usual handful. Much can be said in favor of a large number of participating laboratories. Information regarding the prevalence of systematic errors can be obtained only by having enough laboratories to reveal them and to estimate fairly, by their consensus, the systematic error of the procedure. There is another easy way to enlarge the number of points. An additional pair of different materials, still rather similar to the first pair, are sent to the same laboratories. The results are used to prepare a second graph. The second graph is placed on the top of the first graph, so that the horizontal and vertical lines are coincident and all the points transferred to one graph. This merely gives a common consensus point. As the true compositions have not been used, the absolute values are not involved. If the true compositions are known, the common graph is prepared by plotting the true point on each graph and superimposing these points. The axes are kept parallel. Laboratory numbers should be attached to the points. If a laboratory has both its points far out along the 45° line, the conclusion is obvious to all concerned.

The whole process should be repeated with materials having very different per cent values of the element to be determined. A separate estimate of the precision is proper and should be made. Indeed the systematic error of the procedure may change and possibly that of individual laboratories. The range of per cent and types of materials that require study depend on the analytical chemistry involved.

Discussion

The economy of effort achieved by the elimination of duplicates and other ramifications such as an elaborate schedule of operators, days, etc., is considerable. More important, the rather spurious yardstick of parallel duplicates by the

same operator is discarded. Parallel duplicates are favored indeed. Whatever the attendant circumstances, these duplicates have everything in their favor as far as showing agreement is concerned. Just what use can be made of such a yardstick? Nearly every practical comparison involves determinations carried out under less uniform conditions than a pair of parallel duplicates. Even the single analyses on the two materials are likely to be run together, so that there is the same criticism to be directed against using these to estimate precision. The two materials would be better run at least on different days. Figure 2 shows a plot of potassium determinations by 14 laboratories on two samples of fertilizer. The two samples were run a month apart, so that the estimate of precision is realistic. The clear evidence of individual systematic errors in materials run a month apart shows the persistence of systematic errors.

The estimate of precision proposed here is usually optimistic. A laboratory runs two materials, no doubt under parallel conditions. The two results provide an estimate of the *difference* between the two materials. When the difference is taken between the two results, any common effects drop out, so that the difference is in large measure freed of any consequences of the particular set of circumstances existing when this pair of determinations was made. Every laboratory provides an estimate of the difference and the estimate of the precision is based upon the concordance of these several estimates of the difference between the two compounds. The most that can be said in support of this scheme is that, unlike duplicates on one material, the laboratories do not know the difference between the two materials. There is no protection against a laboratory that runs two or more determinations on each material and reports the averages of these under the label that they are single determinations. Eventually, if over a number of times, a given laboratory always has a point unusually close to the 45° line, it might reasonably be asked to disclose how it consistently achieves a precision so much better than other laboratories.

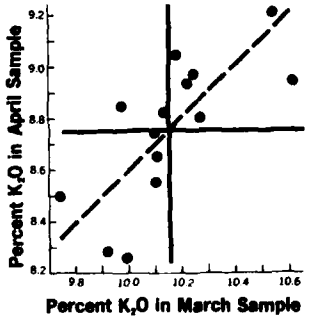


Figure 2

Very careful efforts on analytical work are associated with atomic weight determinations and with the work on standard samples or reference materials. The approach here is chemical rather than statistical. Using every iota of available chemical information elaborate precautions are taken to eliminate, or correct for, every possible source of systematic error. Comparatively little dependence is placed upon repeat determinations. Here the chemist supplies his own testimony to support the position taken in this paper. Systematic errors are the real headache. If enough care is taken, or alternative procedures are employed, the systematic error can be greatly reduced. By such means atomic weights and standard samples gain acceptance. In the ordinary work of analytical chemistry, most of these precautions are not feasible. Nevertheless the goal of general agreement among laboratories, using a procedure with a very small bias, is the task of the analytical laboratories. To achieve their goal, the laboratories must get the right kind of data and interpret them properly.

Literature Cited

(1) Power, F. W., ANAL. CHEM. 11, 660 (1939).
(2) Youden, W. J., *Ind. Eng. Chem.* 50, 83 A (August); 91 A (October); 77 A (December, 1958).
(3) Youden, W. J., *Ind. Quality Control* 15, No. 11, 24 (1959).
(4) Youden, W. J., *Technometrics* 1, 409 (1959).

NBS Miscellaneous Publication 248
Proceedings of the 1962 Standards Laboratory Conference

Session 5. Measurement Agreement Comparisons Among Standardizing Laboratories

Paper 5.1. Measurement Agreement Comparisons

W. J. Youden*

The best source of information on the measurement errors in comparisons is found in the records of the comparisons regularly carried out by a laboratory. This requires that some of the comparisons must be repeated, either directly or indirectly. An item, A, may be compared with a standard, S, and the comparison repeated. Generally it is better to plan the work so that the operator is not directly aware of how well his results check. If two items, A and B, are each compared with S, and then A and B compared directly, the additional measurement provides a check on the measurement process. Thus in addition to the direct comparison of A with S there is the indirect comparison obtained by adding $(A-B)$ to $(B-S)$. The sum of these two comparisons would check the direct result exactly if the measurements could be made without error. The discrepancy between $(A-S)$ and $(A-B)+(B-S)$ must arise from measurement error. Information collected over a sequence of such triads soon provides a sound basis for evaluating measurement error. This example and similar ones will be discussed in some detail in the paper.

1. Introduction

A calibrating laboratory must have in its possession appropriate standards with values certified by a competent authority. The calibrating laboratory must also possess adequate facilities for comparing its standards with items brought to it for calibration. The first thing the calibrating laboratory must attend to is to determine the accuracy of these comparisons. There are other problems such as the appropriate way to combine the comparison error with the uncertainty in the value assigned to the standard. This problem, incidentally, is only important when the comparison error is nearly as small as the uncertainty in the standard. This discussion is concerned with methods for ascertaining the accuracy of the comparisons, and also with getting the most information out of the measurements actually made.

2. Determination of the Accuracy of a Comparison Procedure

2.1. Two Independent Systems for Comparisons

It is not generally possible to attain absolute accuracy. Even if the calibrating laboratory has two similar certified standards and two completely independent assemblies for making comparisons, it is practically certain that, if enough items are calibrated with each of the two independent systems, a difference between the two systems can be demonstrated. This difference may be of negligible importance but once shown to exist, this difference is a component in the absolute error. Even when the calibrating laboratory shows this difference to be extremely small, there is the troublesome thought that the source certifying the two standards may have had some unknown error which was carried over into both certifications.

Such a series of duplicate tests with two independent systems on a succession of items furnishes the data for determining the accuracy of the comparison procedure.

*Consultant, Applied Mathematics Division, National Bureau of Standards, Washington, D.C.

Table 1. Data from two independent calibration systems

Item No.	A	B	C . .	N
System 1	a_1	b_1	c_1 . .	n_1
System 2	a_2	b_2	c_2 . .	n_2
Difference	D_1	D_2	D_3 . .	D_n

Examination of data tabulated as above should reveal whether the D's tend to be predominantly of one sign. The signs of the D's should, if the systems are equivalent, alternate in a random manner. The variance of the comparison process is estimated by calculating

$$s^2 = \frac{\Sigma D^2 - (\Sigma D)^2/n}{2(n-1)}$$

The square root of s^2 gives the standard deviation. This standard deviation (a measure of the precision) applies to any difference, Δ, found between a standard and a test item. It is this difference that applied to the certified value of the standard gives the value entered in table 1.

If the algebraic average for D is unacceptably large, this implies some persistent difference in the two systems. The obvious thing to do is to interchange the two standards with the two sets of comparison equipment. A further series of results will establish whether the discrepancy between the two systems arises from an inconsistency of the two standards or some lack of equivalence in the two sets of comparison equipment. Should the latter be the case, a suitable swapping back and forth of components of the systems will track down the source of inaccuracy in the comparison procedure [1].[1]

2.2 One System With One Standard

The usual technique for ascertaining the error in a comparison procedure is to repeat some of the measurements. This technique has the virtue of simplicity but it may not be the best way of obtaining data to determine the error in the comparison procedure. Direct repetition is vulnerable to repeating the same misreading of a scale. It is also vulnerable to "memory" or operator efforts to secure good checks. Few can resist the temptation, if a pair of results differs rather more than usual, to do one of two things--(a) To reject the pair of results and repeat the readings, or (b) To take a third reading and pair it with the closer of the first two readings. Many operators are unaware that if the average absolute difference between duplicate readings is R then about 11 percent of the individual differences legitimately exceed $2R$. If differences are rejected solely because they slightly exceed twice the average difference, the 'average' difference gradually becomes smaller. More stringent rejection will further reduce the average of the survivors. The logical end of this process is apparent reduction of the error to zero but at the price of rejecting all of the measurements. Another shortcoming of direct repetition is that there are alternatives that are slightly more efficient in estimating Δ, the difference between the standard and the item to be calibrated. More important, these alternatives reduce the number of times the standard is used and thus cut down on any wear or other consequences that follow from repeated use of a standard.

Quite commonly meter bar calibrations included not only comparisons of the standard with each bar but all possible comparisons among the bars in the group. Recently [2] the use of selected subsets of the pairings have been found satisfactory. On the other hand studies with standard cells tend to repetitive comparisons of a standard with the other cells and to make little if any use of intercomparisons among the cells. It seems likely that use would be found for schemes that replace most, if not all, of the repeat measurements by intercomparisons among a group of items only some of which are ever directly matched against the standard. This technique assumes that the test items are similar to and of comparable quality to the standard and also that the environmental control for the test items is equivalent to that maintained for the standard.

The principle of such schemes is shown by the example of comparing two items, A and B, with a standard S. We will suppose that the comparisons $(S\text{-}A)$ and $(S\text{-}B)$ are each repeated three times as is often done. Each set of three results provides an estimate of the variance with two degrees of freedom so the work provides a total of four degrees of freedom. A series of such sets of data will build up the number of degrees of freedom to give a better estimate of the variance. Note that the average of the three measurements of the difference between standard and test item has one third the variance of a single measurement.

A suggested scheme compares S with A, S with B, and A with B. Each comparison is repeated once. Observe that even if S and B were not directly compared, an estimate of $(S\text{-}B)$ is available by adding to $(S\text{-}A)$ the result for $(A\text{-}B)$. This information on $(S\text{-}B)$ can be averaged with the direct comparison of S and B. More weight is given the direct comparison. In this case the theory of least squares gives the direct comparison twice the weight of the indirect comparison. Denote $(S\text{-}A)$ by a, $(S\text{-}B)$ by b and $(A\text{-}B)$ by c. The weighted average for $(S\text{-}B)$ is given by $(2b+a+c)/3$. Similarly the weighted average for $(S\text{-}A)$ is given by $(2a+b-c)/3$. The variance for the average difference between standard and item when each of the three comparisons has been measured twice is again one third of the variance of a single measurement. Three degrees of freedom for error come from the three pairs and a fourth degree of freedom from the fact that $(S\text{-}A)+(A\text{-}B)+(B\text{-}S)$ should be zero in the absence of error of measurement. Consequently $(a+c-b)^2/3$ should be added to the sum of the squared differences of the duplicate readings. The square root of one fourth of this total gives s. If $(a+c-b)^2/3$ tends to be generally larger than the squared differences from duplicates, there is evidence of a certain amount of "forced" agreement between the duplicates. The scheme cuts the use of the standard by one third, retains the same variance for comparisons, and provides a check on the technique of measurement.

A scheme for three items (fig. 1, Scheme II) avoids the repetition of any measurement and cuts

[1] Figures in brackets indicate the literature references at the end of this paper.

the use of the standard in half. All possible six pairs of S, A, B, and C are compared. The average for $(S-A)$ is computed by combining five of the measured differences as follows

$$1/4\,[2(S-A)+(S-B)+(S-C)+(B-A)+(C-A)].$$

From symmetry, averages for all six comparisons are easily obtained. The six discrepancies between these calculated averages and the matching direct measurements reveal the measurement error. These discrepancies tend to be smaller than the differences between duplicates. The six discrepancies are squared. One third the sum of the six squares gives the variance of a single comparison. The variance of the average difference between standard and item is half that of a single comparison--just what the duplicate readings would give.

When there are four test items Scheme III, instead of duplicating the comparisons $(S-A)$, $(S-B)$, $(S-C)$, $(S-D)$, calls for comparisons $(A-B)$, $(B-C)$, $(C-D)$, and $(D-A)$. Now the calculated average for $(S-A)$ has a variance of 7/15 of a single comparison which is a small improvement over the 1/2 that simple duplication would give.

Scheme IV (fig. 1) reduces the use of the standard over Scheme III and provides for more information on some test items than on others. There are times when this discrimination among items is convenient. Scheme V reduces both the use of the standard and the number of measurements and hence reduces the degrees of freedom available for the variance estimate. In a continuing program this reduction in the amount of duplication may be acceptable if duplication is used largely to maintain a check on operations. Scheme V, interestingly enough, provides equal information on all four test items in spite of the corner position for the standard.

Schemes II, III, and VI are the first three of a series formed in a particular way. Beginning with Scheme III the comparison between standard and item has a smaller variance (about 7%) than straight duplicates would provide. The feedback through the comparison links brings about this improvement in efficiency.

Schemes VII, VIII, and IX show some additional patterns that may be extended to larger numbers of items. The tenth pattern illustrates a scheme making use of two standards. Clearly a wide variety of schemes can be devised. This permits the laboratory to select schemes appropriate for its particular program.

Two illustrative numerical examples are included. Formulas are not given for each scheme shown because they may be obtained from a statistician or a least square fit made to the data. A short cut for determining the weighting coefficients for the observed quantities is based upon an analogy with an electric circuit. The lines in the diagrams may be considered as one ohm resistances. If a potential is maintained between any two points the resulting equilibrium currents in the network give the relative weighing coefficients for the observations used to estimate the measurement comparison between the quantities represented by the two points to which the potential has been applied. Thus, in Scheme VII, if a potential of 3 v is applied between the standard and the midpoint of any side the current flow in the various resistances are exactly those shown in the first three lines of the illustrative example. A more detailed discussion is under preparation.

3. Interlaboratory Comparisons

It is common practice to send a "package" of several similar items on a circuit of several laboratories. The data should be examined to see if there is evidence that a particular laboratory tends to report consistently higher (or lower) values than the other participating laboratories.

One method of statistical analysis consists in taking the data for one of the items and assigning the rank of one to the laboratory with the highest value, the rank two to the laboratory with the next highest value, and so on. If there are L laboratories, the laboratory with the lowest value receives the rank L. This ranking procedure is carried out for each of the M items included in the package. A "score" for each laboratory is obtained by adding up the M ranks assigned to each laboratory. If a laboratory tends to get high values, its score will be low, but not lower than M. Low values lead to a high score with a maximum possible score of ML. If only random errors are responsible for the assigned ranks, the expected score is midway between M and ML or $L(M+1)/2$. Scores that depart sufficiently from the expected score constitute evidence of the presence of systematic errors. The attached table 2 shows scores which, if attained, constitute evidence of a systematic error. A detailed account of this new technique is available in Materials Research & Standards. [3]

Table 2

Let L laboratories test each of M materials. Assign ranks 1 to L for each material. Sum the ranks to get the score for each laboratory. The mean score is M(L+1)/2. The entries are lower and upper limits that are included in the approximate 5 percent critical region.

Approximate 5 percent two-tail limits for ranking scores

No. of Labs.	Number of materials												
	3	4	5	6	7	8	9	10	11	12	13	14	15
3		4 12	5 15	7 17	8 20	10 22	12 24	13 27	15 29	17 31	19 33	20 36	22 38
4		4 16	6 19	8 22	10 25	12 28	14 31	16 34	18 37	20 40	22 43	24 46	26 49
5		5 19	7 23	9 27	11 31	13 35	16 38	18 42	21 45	23 49	26 52	28 56	31 59
6	3 18	5 23	7 28	10 32	12 37	15 41	18 45	21 49	23 54	26 58	29 62	32 66	35 70
7	3 21	5 27	8 32	11 37	14 42	17 47	20 52	23 57	26 62	29 67	32 72	36 76	39 81
8	3 24	6 30	9 36	12 42	15 48	18 54	22 59	25 65	29 70	32 76	36 81	39 87	43 92
9	3 27	6 34	9 41	13 47	16 53	20 60	24 66	27 73	31 79	35 85	39 91	43 97	47 103
10	4 29	7 37	10 45	14 52	17 60	21 67	26 73	30 80	34 87	38 94	43 100	47 107	51 114
11	4 32	7 41	11 49	15 57	19 65	23 73	27 81	32 88	36 96	41 103	46 110	51 117	55 125
12	4 35	7 45	11 54	15 63	20 71	24 80	29 88	34 96	39 104	44 112	49 120	54 128	59 136
13	4 38	8 48	12 58	16 68	21 77	26 86	31 95	36 104	42 112	47 121	52 130	58 138	63 147
14	4 41	8 52	12 63	17 73	22 83	27 93	33 102	38 112	44 121	50 130	56 139	61 149	67 158
15	4 44	8 56	13 67	18 78	23 89	29 99	35 109	41 119	47 129	53 139	59 149	65 159	71 169

4. Summary

Calibration requires measuring the difference between a standard and a test item. Systematic errors can, with care, be practically eliminated from comparisons. Repeat measurements are generally used to estimate the precision of the comparisons. Repeat measurements may not be as independent as they should be. This paper lists various schemes that replace repeat determinations by comparisons among the test items. The advantages are (i) reduced use and wear on the standard; (ii) a more valid estimate of the precision; (iii) a slight improvement in the information obtained from a given number of measurements; and (iv) a flexible program adaptable to various programs.

A brief description of a new ranking procedure useful in interlaboratory tests is given together with a table.

5. References

[1] Youden, W. J., Experimental design and ASTM Committees, Mater. Res. Std. **1**, 862-867 (1961).

[2] Page, B. L., Calibration of meter line standards of length at the National Bureau of Standards, J. Research **54**, 1-14 (1955), RP2559.

[3] Youden, W. J., Ranking Laboratories by Round-Robin Tests, Mater. Res. Std. **3**, No. 1, 9-13 (1963).

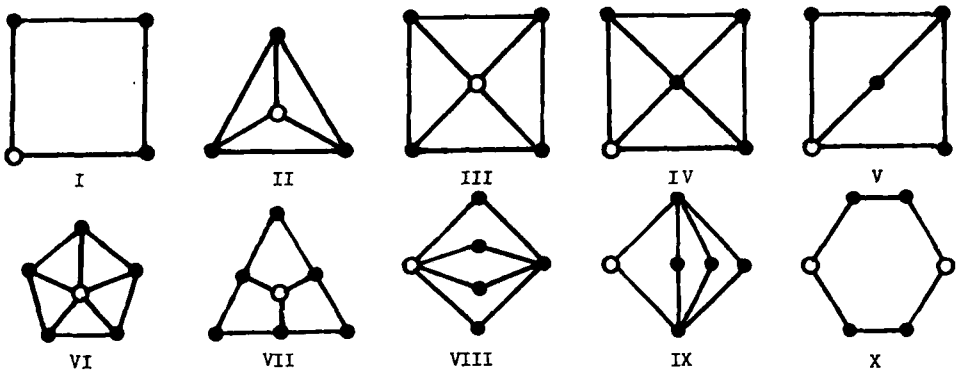


Figure 1. Calibration schemes. Circles identify the standard, solid dots represent test items, and connecting lines show the comparisons that are measured.

Illustrative examples:

Scheme I with data obtained in transposition of 10 gram weights.

Measured: S-A=-.011; S-B=.068; A-C=-.023; B-C=-.105 (mg.)

Calculated: $S\text{-}A = \frac{1}{4}[3(S-A)+(S-B)-(A-C)+(B-C)] = -.01175$; $S\text{-}B = .06875$

$A\text{-}C = -.02375$; $B\text{-}C = -.10425$

$S\text{-}C = \frac{1}{2}[(S-A)+(A-C)+(S-B)+(B-C)] = -.0355$

Variance = Σ (diff. between measured and cal.)^2 = .00000225

s = .0015

Scheme VII using data taken with meter bars. See reference 2.

Nine pairings taken from a ten bar study using all 45 pairings.

Bar identifications: S=27; A=4; B=21; C=39; D=153R; E=752; F=814B

Pair observed	Measured value	Multiplying coefficients a	b	c	d	e	f	g	h	i	Divide by	Cal. value	Obs.-cal.
S-A	a= 4.33	3	1	1	-1	1	0	0	1	-1	5	4.272	.058
S-B	b= -5.11	1	3	1	1	-1	-1	1	0	0	5	-5.090	-.020
S-C	c=177.13	1	1	3	0	0	1	-1	-1	1	5	177.168	-.038
A-D	d= 19.50	-2	2	0	7	3	-1	1	-1	1	10	19.469	.031
B-D	e= 28.80	2	-2	0	3	7	1	-1	1	-1	10	28.831	-.031
B-E	f=184.94	0	-2	2	-1	1	7	3	-1	1	10	184.929	.011
C-E	g= 2.66	0	2	-2	1	-1	3	7	1	-1	10	2.671	-.011
C-F	h= -6.96	2	0	-2	-1	1	-1	1	7	3	10	-6.933	-.027
A-F	i=165.99	-2	0	2	1	-1	1	-1	3	7	10	165.963	.027
S-D	(23.70)*	4	4	2	5	5	-1	1	1	-1	10	23.741	---
S-E	(179.80)*	2	4	4	1	-1	5	5	-1	1	10	179.839	---
S-F	(170.34)*	4	2	4	-1	1	1	-1	5	5	10	170.235	---

$\Sigma(\text{Obs.}-\text{cal.})^2 = 0.008830$; Stand. Dev. = $\sqrt{.008830/3} = 0.054$

*Measured by Page. Not used in these calculations to estimate S-D, S-E, and S-F.

The Collaborative Test*

By W. J. YOUDEN (National Bureau of Standards, Washington, D.C.)

This paper discusses (a) the planning of collaborative tests, (b) a technique to establish that a procedure is ready for a collaborative test, and (c) the interpretation of the results of a collaborative test.

Introduction

The collaborative, or interlaboratory, test is an indispensable scrutiny of an analytical procedure to insure (a) that the description of the procedure is clear and complete and (b) that the procedure does give results that are in accord with any accuracy claims made for the procedure. A collaborative test should be a kind of final inspection. If the procedure has been properly studied before submitting it to a collaborative test, then the collaborative test has as its proper role the task of verifying any claims made for the procedure.

Planning a Collaborative Test

There are three matters to settle in planning a collaborative test. These are the number of collaborators, the number of materials sent to each collaborator, and the number of measurements made by each collaborator on each material. Inevitably certain compromises have to be made. A large number of collaborators is desirable because this will give confidence that analysts will not misinterpret the instructions and that the procedure has been tried under a wide range of environments. Increasing the number of materials provides evidence that the procedure is satisfactory over a wide range of amounts present and types of material. Repeat analyses on each material would provide information on the agreement of parallel analyses made under as nearly identical conditions as possible.

Increasing the number of materials and the number of analyses on each material adds considerably to the burden of work imposed on each collaborator. Often this has the unfortunate consequence of reducing the number of laboratories willing to participate as collaborators. Therefore it is important to hold to a minimum the work imposed on each collaborator. One only has to consider two extreme situations to see the importance of having an adequate number of collaborators. If you want to learn about a procedure, which would you rather have: Ten repeat analyses from one laboratory or a single analysis from each of ten laboratories? True, the information given by these alternatives is quite different, but the really useful information is given by the single results from the ten laboratories.

The best way to reduce the workload per laboratory is to reduce the number of repeat analyses made on each material (1). In spite of the long tradition to require at least duplicate determinations on each material, a strong case can be made for requiring just single determination per material, unless repetitions are actually needed. There are several reasons behind this suggestion. First, the agreement of parallel determinations should be about as good in one laboratory as in another. After all, the equipment is specified and there is the presumption of qualified analysts. Certainly the laboratory environment will vary from laboratory to laboratory and the procedure may not be immune to these changes in environment. But *within* any one laboratory, parallel determinations will be exposed to the same environment and the agreement between the duplicates normally will not be impaired by reason of any local environmental peculiarity. For this reason it is not surprising that

* Presented at the Referees' Meeting, Seventy-sixth Annual Meeting of the Association of Official Agricultural Chemists, Oct. 16, 1962, at Washington, D.C.

the *precision,* as revealed by repeat runs, is indistinguishably the same for all participating laboratories.

A second reason for not requiring repeat determinations is that rarely are enough data available to detect a two-fold difference in precision (standard deviation) between two laboratories. Triplicate determinations on each of seven materials will give a four out of five chance of catching a two-fold difference in precision. It would take the equivalent of five repeat determinations on each of ten materials to have the same probability of detecting that one laboratory has a standard deviation 1.5 times that of another laboratory. Clearly this is a lot of extra work for each laboratory. On the other hand, the initiating laboratory should have ample records to establish the precision of the procedure. The precision, in any event, is usually of minor importance as compared with the larger error inevitably associated with the comparison of results from different laboratories.

One might also mention that many laboratories will not report a pair of duplicates that happen to show rather poor agreement. The temptation to run a third determination, or even another pair, is strong. The consequences of any such censoring of the data is to produce an estimate of the precision that is biased in the direction of making the precision appear to be better than it really is. Finally the precision can be estimated even if only single determinations are made, and such an estimate is immune from any replacements of the results first obtained. It is merely necessary that two materials, A and B, similar in composition and content, be included in the work. Let the results from n laboratories be as follows:

Material	*Laboratory Number*					*Av.*
	1	2	3	....	n	
A	a_1	a_2	a_3	....	a_n	$\bar{a}$
B	b_1	b_2	b_3	....	b_n	$\bar{b}$
Difference (A − B)	d_1	d_2	d_3	....	d_n	$\bar{d}$

Compute $s = \sqrt{\frac{\Sigma d^2 - n\bar{d}^2}{2(n-1)}} =$ standard deviation

You will observe that whatever local or systematic error a laboratory has drops out of the differences, $d_1, d_2, \ldots, d_n$. These differences should all be the same except for precision errors. So it is the variation among these differences that provides an estimate of the precision. The above formula is equivalent to deducting the mean difference, $\bar{d}$, from each of the n differences and calling the remainders d′. Thus $d_1 - \bar{d} = d_1'$. These remainders are squared and divided by $2(n-1)$, and the square root is taken.

$$s = \sqrt{\Sigma(d')^2/2(n-1)}$$

An estimate of the precision by this approach is more realistic in that it is protected against any selection of the data by replacement of repeat determinations that show larger than usual disagreement and the estimate is a consensus taken over all the participating laboratories.

We arrive, then, at the suggestion that the collaborative test include as many laboratories as possible, using as many materials as circumstances suggest, and that only single determinations be required. Some have raised the question that certain laboratories might run duplicates but report the averages as single determinations. A laboratory that does this is ill advised. First, the averages of two would give this laboratory an apparent standard deviation of only 0.707 that of laboratories running single determinations. But the data will not visibly reveal this if only because of the difficulty of showing small differences in precision. Rather less pleasing to such a laboratory is that this average reveals only the more clearly any systematic error the laboratory has in comparison with the consensus of all the laboratories. And it is on just this point that attention is going to be focused with the idea of asking such laboratories for explanations.

The Responsibility of the Initiating Laboratory

By no means an unusual occurrence is a collaborative test whose results obviously fall short of expectations based on data obtained by the initiating laboratory. The explanation is usually found in the fact that

the initiating laboratory has a set of operations and equipment that is never varied. In fact, care is taken not to vary the routine in any particular. Naturally no light is shed on what may happen when the procedure on trial is used by a number of laboratories each of which establishes its own particular routine. Such things as the source and age of reagents and the concentrations of these reagents, the rate of heating, thermometer errors, humidity, and many other factors may be involved. One laboratory makes up a supply of nominally 1*M* acid and in fact achieves a concentration of 0.95. Another laboratory's solution may be 1.03*M*. Each laboratory gets good checks, of course, because it always uses the same solution, just as the initiating laboratory did.

The only protection against such sources of trouble which are disconcerting and difficult to discover is for the initiating laboratory deliberately to introduce minor reasonable variations in the procedure and observe what happens. These departures should be of the magnitude that a chemist might well expect to find among laboratories. At first this appears to throw much extra work on the initiating laboratory, but if the program is carefully laid out, a surprisingly small amount of work suffices.

We will suppose that as many as seven factors are selected for scrutiny. Perhaps the volume of solution is fixed at 100 and 110 ml; the time of waiting at some stage is tried at 30 and at 40 minutes. Different lots of reagent, slightly different concentrations, different times to bring solutions to boiling may also be tried. Now, if the procedure is "rugged" and therefore immune to modest (and inevitable) departures from some habitual routine, the results obtained should not be altered by these minor departures. If the results are altered, we should by all means know about it and warn the prospective user not to depart by more than some stated amount from the specified condition. Presumably most of these minor departures will show negligible effects, but if just one sensitive condition is spotted, we may save the very considerable effort that would have been expended in a disappointing collaborative test—particularly disappointing because it is all but impossible to track down the responsible conditions, since all the laboratories quite sincerely report that they followed the procedure.

What is needed is a scheme of attack that will conserve labor yet be sensitive enough to pick up fairly small effects if they should occur when some condition has been slightly altered. Negligible effects will be found for most changes. There is a program for making slight modifications in the procedure that has a very high efficiency in identifying those changes that do produce effects. The basic idea is not to study one alteration at a time but to introduce several changes at once, in such a manner that the effects of individual changes can be ascertained. Let A, B, C, D, E, F, and G denote the nominal values for seven different factors that might influence the result if their nominal values are slightly changed. Let their alternative values be denoted by the corresponding lower case letters a, b, c, d, e, f, and g. Now the conditions for running a determination will be completely specified by writing down these seven letters, each letter being either a capital or lower case. There are 2^7 or 128 different combinations that might be written out. Fortunately it is possible to choose a subset of eight of these combinations that have an elegant balance between capital and lower case letters.

The particular set of combinations is shown in Table 1. The table specifies the values for the seven factors to be used while running eight determinations. The results for the analyses are designated by the letters *s* through *z*. Let us see how to extricate the separate effects of the factor changes, even though four factors are always altered from the initial combination of all capitals. To find whether changing factor *A* to *a* had an effect, we compare the average $(s + t + u + v)/4$ with the average $(w + x + y + z)/4$. The table shows that determinations 1, 2, 3, and 4 were run with the factor at level *A* and determinations 5, 6, 7, and 8 with the factor at level *a*. Observe that this partition gives two groups of four determinations and that each group contains the other six factors twice at the capital level and twice at the lower case

Table 1. Eight combinations of seven factors used to test the ruggedness of an analytical procedure

Factor Value	Combination or Determination Number							
	1	2	3	4	5	6	7	8
A or a	A	A	A	A	a	a	a	a
B or b	B	B	b	b	B	B	b	b
C or c	C	c	C	c	C	c	C	c
D or d	D	D	d	d	d	d	D	D
E or e	E	e	E	e	e	E	e	E
F or f	F	f	f	F	F	f	f	F
G or g	G	g	g	G	g	G	G	g
Observed result	s	t	u	v	w	x	y	z

level. The effects of these factors, if present, consequently cancel out, leaving only the effect of changing *A* to *a*.

Inspection of Table 1 shows that whenever the eight determinations are split into two groups of four on the basis of one of the letters, all the other factors cancel out within each group. Every one of the factors is evaluated by all eight determinations. The effect of altering *G* to *g*, for example, is examined by comparing the average $(s + v + x + y)/4$ with the average of $(t + u + w + z)/4$. Suppose only six factors are explored. In that event, associate with *g* some meaningless operation such as solemnly picking up the beaker, looking at it intently, and setting it down again. Omit this meaningless operation for the determinations that involve *G*. (Be sure to look at the average difference between the G's and g's, because if they are large an explanation should be sought!)

Collect the seven differences for A − a, B − b, . . . , G − g, and list them in order of size. If one or two factors are having an effect, their differences will be substantially larger than the group of differences associated with the other factors. Indeed, this ranking is a direct guide to the procedure's sensitivity to modest alterations in the factors. Obviously a useful procedure should not be affected by changes that will almost certainly be encountered between laboratories. If there is no outstanding difference, the most realistic measure of the analytical error is given by the seven differences obtained from the averages for capitals minus the averages for corresponding lower-case letters. Denote these seven differences by D_a, D_b, . . . , D_g. To estimate the standard deviation, square the differences and take the square root of 2/7 the sum of their squares. This estimate of the analytical error is realistic in that the sort of variation in operating conditions that will be encountered among several laboratories has been purposely created within the initiating laboratory. If the standard deviation so found is unsatisfactorily large, it is a foregone conclusion that the collaborative test will also give disappointing results. The collaborative test should never be undertaken until a procedure has been subjected to the abuse described above and satisfactory results obtained in spite of the abuse.

The schedule shown in Table 1 can be modified in various ways. An interesting variant is to replace the capitals with lowercase letters and vice versa. This creates eight new combinations. If all sixteen combinations are tried, smaller effects will be detected as well as possible mutual interferences of the factors. At this point a statistician will likely be of considerable assistance. There will be some who may see in this scheme a means of studying a procedure in its formative stage. Generally this is inadvisable, because substantial changes in the factors seldom act independently and a more complex schedule of factor values is appropriate. There are also schedules for eleven and fifteen factors which may be found useful (2–5).

If only those procedures that survive this

planned introduction of minor modifications in the procedure were submitted to a collaborative test, then the latter would really take on the role of confirming that a good procedure has in fact been devised. Much disappointment would be avoided and sources of difficulty would be tracked down by this planned work within one laboratory. It should not be necessary to involve several laboratories in order to discover serious shortcomings in a procedure. Fewer collaborative tests would be needed and participation would be encouraged because the chance of a successful outcome would be very high.

The Interpretation of the Data

After the cooperating laboratories have made their reports, the results may be tabulated as shown in Table 2. Generally one would hope for a table with about forty or more entries, and every effort should be made to avoid missing entries.

It is useful to consider Table 2 as a whole and try to place the table in one of four categories. The hoped-for category is that the standard deviation as calculated for each column in the above table is acceptably small.

If $x_1, x_2, \ldots, x_n$ are the results tabulated in a column for any one material, the estimate for the standard deviation for that column is given by

$$s = \sqrt{(\Sigma x^2 - n\bar{x}^2)/(n-1)},$$

where $\bar{x}$ is the mean for the column. The standard deviation may, of course, vary with the amount present and it would be informative to prepare a graph plotting the standard deviation as ordinate against the amount present as abscissa. Some irregularity is to be expected, particularly if fewer than ten laboratories participate. A smooth curve should be drawn in with no attempt to follow the individual ups and downs. Values of the standard deviation read from this curve are very likely closer to the mark than the individual points. If the curve is approximately a straight line going near the origin, then the error is proportional to the amount present. Very often, in such an event, the error is expressed as per cent of the amount present and labeled the "Coefficient of Variation."

If the standard deviation when plotted against the amount present gives a series of points that show no trend, then the best fit is a horizontal line $Y = s^*$. That is, the standard deviation is the same over the range of amount present used in the work. The best value to use for s^* is not the average of the standard deviations found for the "M" columns. The squares of the standard deviations should be summed and divided by M, and the square root taken to get the best estimate of the standard deviation that will be appropriate for all the materials. This estimate of the standard deviation has $M(n-1)$ degrees of freedom. There should be at least 20 degrees of freedom to provide a reasonably good estimate of the standard deviation.

If the standard deviation as calculated for all, or most, of the columns is unacceptably large, the table of data may usually be classed in one of three categories. In order

Table 2. Tabulation of results

Laboratory No.	Material Number						
	A	B	C	D	—	—	M
1	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—
n	—	—	—	—	—	—	—

to determine the category, a convenient device is to prepare another table that better reveals certain features of the data, as follows: Scan the entries in the first column of Table 2 and assign the rank of 1 to the highest result, the rank of 2 to the next highest result and so on, until the rank of n is given to the lowest result in that column. Enter these ranks opposite the appropriate laboratories in the first column of the new table. If two laboratories are tied for fourth place, assign to each the rank of 4.5. If three are tied for second place, assign all three the rank of 3. This keeps the sum of the ranks, $n(n+1)/2$, the same for each column. Repeat this process for each column, and then sum the ranks assigned to the first laboratory and enter it as a laboratory score at the right of the row. Sum the ranks in each row. When the scores achieved by all the laboratories are added, the total should be $Mn(n+1)/2$, and this provides a convenient check on the work.

Should a laboratory turn in the highest result for each of the "M" materials, its score would be M, the lowest possible. The highest possible score is nM and the average score is $M(n+1)/2$. The scores obtained by the n laboratories afford certain clues as to the reason why an unsatisfactory standard deviation was obtained from the reported results. The interpretation depends on the fact that for each combination of n laboratories and M materials, it is possible to compute a lower and an upper limiting score. Scores as low as or lower, or as large as or larger than these limiting scores are an indication of trouble. They mean that a laboratory with such an extreme score has a definite tendency to get persistently high or low results.

Now it is possible for the standard deviation to be unacceptably large and yet for no laboratory to turn up with an extremely low or high score. This would happen if the precision of the method is very poor. It may also happen if a laboratory tends to get high or low results for materials with low percentages and opposite results with materials of high percentages. If this happens with several laboratories, scores tend to cluster near the average score. Whatever the explanation, the evidence points to some defect in the procedure.

Another category arises when one or perhaps two laboratories have quite extreme scores. This laboratory (or both, if there are two) is the one chiefly responsible for the large standard deviations found for the individual columns. If the results from this laboratory are set aside, the standard deviation calculated by using the remaining laboratories may be acceptable. The basis for setting aside these results is that the limiting scores have been so chosen that only one collaborative test in twenty can be expected to include an extreme score by chance. An extreme score is, in consequence, a strong hint that the laboratory concerned has a pronounced bias, probably as a result of

Table 3. Water-insoluble nitrogen results

Coll. No.	Results, %					Ranked Results					Coll. Score
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	
7	4.59	1.46	5.64	2.19	27.32	9	5.5	6	4	3	27.5
8	4.94	1.52	5.68	2.28	26.44	1	1	3	2	10	17
9	4.80	1.40	5.62	2.12	26.89	3.5	8.5	7.5	6.5	8	34
10	4.73	1.46	5.65	2.09	27.17	5	5.5	5	8	4	27.5
11	4.72	1.51	5.62	2.12	27.00	6.5	2.5	7.5	6.5	6	29
12	4.80	1.51	5.80	3.29	27.48	3.5	2.5	1	1	1	9[a]
13	4.45	1.40	5.45	2.07	27.02	10	8.5	10	9	5	42.5
15	4.72	1.50	5.58	2.27	26.76	6.5	4	9	3	9	31.5
16	4.63	1.32	5.69	2.04	26.92	8	10	2	10	7	37
17	4.88	1.42	5.67	2.16	27.39	2	7	4	5	2	20

[a] Designates unusually low score.

Let n laboratories test each of M materials. Assign ranks 1 to n for each material. Sum the ranks to get the score for each laboratory. The mean score is $M(n+1)/2$. The entries are lower and upper limits that are included in the approximate 5% critical region.

Table 4. Approximate 5% two-tail limits for ranking scores

No. of Labs.	Number of Materials												
	3	4	5	6	7	8	9	10	11	12	13	14	15
3		4	5	7	8	10	12	13	15	17	19	20	22
		12	15	17	20	22	24	27	29	31	33	36	38
4		4	6	8	10	12	14	16	18	20	22	24	26
		16	19	22	25	28	31	34	37	40	43	46	49
5		5	7	9	11	13	16	18	21	23	26	28	31
		19	23	27	31	35	38	42	45	49	52	56	59
6	3	5	7	10	12	15	18	21	23	26	29	32	35
	18	23	28	32	37	41	45	49	54	58	62	66	70
7	3	5	8	11	14	17	20	23	26	29	32	36	39
	21	27	32	37	42	47	52	57	62	67	72	76	81
8	3	6	9	12	15	18	22	25	29	32	36	39	43
	24	30	36	42	48	54	59	65	70	76	81	87	92
9	3	6	9	13	16	20	24	27	31	35	39	43	47
	27	34	41	47	54	60	66	73	79	85	91	97	103
10	4	7	10	14	17	21	26	30	34	38	43	47	51
	29	37	45	52	60	67	73	80	87	94	100	107	114
11	4	7	11	15	19	23	27	32	36	41	46	51	55
	32	41	49	57	65	73	81	88	96	103	110	117	125
12	4	7	11	15	20	24	29	34	39	44	49	54	59
	35	45	54	63	71	80	88	96	104	112	120	128	136
13	4	8	12	16	21	26	31	36	42	47	52	58	63
	38	48	58	68	77	86	95	104	112	121	130	138	147
14	4	8	12	17	22	27	33	38	44	50	56	61	67
	41	52	63	73	83	93	102	112	121	130	139	149	158
15	4	8	13	18	23	29	35	41	47	53	59	65	71
	44	56	67	78	89	99	109	119	129	139	149	159	169

deviation, unintentional or otherwise, e procedure.

is point it appears proper to query laboratory with an extreme score to rtain if the laboratory can offer any anation for its results being consistently higher (or lower) than the results of the other participants.

In a very real sense a collaborative test reveals not only the performance of the procedure under test but also the performance of the laboratories doing the testing. The

intent of this ranking device is to prevent a procedure from being unjustly rated poor when one or two laboratories are in fact responsible for the large scatter of the results.

Finally, the last category of unsatisfactory collaborative tests contains clearly unsatisfactory procedures. Sometimes the table of ranks shows little or no change in the assigned ranks as the eye moves from column to column in the table. In other words, a laboratory tends to hold its same rank for all materials. Usually there will be at least one very high and one very low score. What this tells is that each laboratory is doing the same thing very carefully every time. Some minor departure from a specified factor value, or even an arbitrarily chosen value for a factor because none was specified, is seriously influencing the analytical results. Obviously each laboratory is carefully following whatever routine it adopted. Now it is ridiculous to say that all the laboratories are inadequate. It makes better sense to conclude that here is a procedure so very vulnerable that it should never have been submitted to a collaborative test.

Illustrative Example of Ranking Technique

Table 3 shows a portion of a rather extensive collaborative test on nitrogen in fertilizers (6). The data for the water insoluble nitrogen are shown in the left half of Table 3 for ten of the participating laboratories. The right half of the table shows the ranks assigned to the collaborators; the rank of one is given to the highest result and the rank of ten to the lowest result on each sample. It happens that the data are, in fact, averages of duplicates but this does not disturb the ranking technique. The result for Sample 4 by Collaborator 12 looks peculiar but even if the 3 is a misprint for 2 the ranking would not be altered.

The last column of the table shows the scores obtained for each collaborator by adding up the 5 ranks obtained with the 5 samples. The critical 5% probability scores for 10 laboratories and 5 samples are 10 and 45. Collaborator 12 runs persistently high and has a score of 9, which is in the critical region. The evidence indicates that Collaborator 12 has some individual manner of making the determination. Critical scores for as many as 15 collaborators and 15 samples are listed in Table 4 (5).

Discussion and Summary

This paper has considered several important aspects of collaborative test programs. The question of the distribution of the analytical effort is of prime importance. A broad basis for judgment requires enough laboratories and materials to be representative of the users and the materials likely to be submitted for analysis. In order to prevent unduly burdensome programs it is recommended that duplicates be eliminated and reliance placed on the initiating laboratory for information as to the precision of the procedure.

Another very important question concerns the need to make sure that the procedure is really ready for a collaborative test and that it will almost surely pass this final inspection. To that end an efficient and systematic way of disclosing possible weaknesses in the procedure has been presented in detail. The initiating laboratory should present evidence of the performance of the procedure when minor and seemingly inconsequential changes are made.

Finally a method has been described for evaluating unsatisfactory collaborative test results which should be valuable as a guide to determining the probable cause of the unsatisfactory results.

References

(1) Youden, W. J., *This Journal*, **45**, 169 (1962).
(2) Plackett, R. L., and Burman, J. P., *Biometrika*, **33**, 305 (1946).
(3) Yates, F., *Royal Statistical Society*, Supplement, **2**, 181 (1935).
(4) Youden, W. J., *Materials Research and Standards*, **1**, 862 (1961).
(5) Youden, W. J., *ibid.*, **2**, in press (1962).
(6) Davis, H. A., *This Journal*, **42**, 494 (1959).

Reprinted from the *Journal of the Association of Official Agricultural Chemists*, Vol. 46, February 1963.

Experimental Design and ASTM Committees

By W. J. YOUDEN

This paper considers the subject of experimental design from the viewpoint of ASTM committee members concerned with devising and evaluating test procedures. Simple designs are described that make for an efficient approach to the identification of defects in a test procedure. Suggestions are made regarding the supporting evidence that should be offered by an initiating laboratory before conducting an interlaboratory test. A preliminary type of interlaboratory test is advanced as a means of checking on the claims made in behalf of a test procedure.

THERE ARE numerous textbooks available that present a systematic account of the various types of experimental design and the analysis of variance appropriate for each type of design. Why then should anyone undertake to write on experimental design for ASTM committees? The answer appears to lie in the fact that statistical texts organize the material on experimental design from the viewpoint of statistics. There is a need for expositions that emphasize the objectives and problems that confront ASTM committee members. This paper discusses certain problems that arise in the progress of a test procedure from its inception to the status of a standard procedure. It offers an approach to those recurring statistical problems of general concern regardless of the particular material involved.

The Inception of a Test Procedure

A new test procedure, or a modification of an old procedure, begins in a laboratory. The research involved in devising a test procedure calls for expert knowledge of the material to which the procedure will be applied. The test procedure must serve a useful purpose. Usually it evaluates some property of the material that must be known within certain limits. Satisfactory estimates of the properties of material are required for the safe and economical use of materials and for the setting of fair values in the exchange of materials. The initiating laboratory should be able to supply certain information before requesting that a group of laboratories participate in a round-robin evaluation. There are defects in a test procedure that are best ascertained by work *within* one laboratory, and only confusion results if the detection of these defects is attempted using the less sensitive comparisons associated with interlaboratory tests.

Many test procedures are used to predict the performance in use of the material undergoing test. The laboratory or agency proposing a test must bear in mind that the test will be used for prediction purposes. Devising an adequate test procedure is often a major research problem. A round robin reveals the agreement or lack of agreement among the test results obtained by different laboratories. Agreement among the laboratories does not establish that the test procedure provides a satisfactory measure of the performance of the material. Agreement among laboratories is a necessary even though not a sufficient criterion of a good test procedure. This paper takes up the question of attaining agreement among laboratories.

A test is made on a sample or specimen. The initiating laboratory should, based on its expert familiarity with the material, undertake to specify what sort of a sample, or samples, composite or otherwise will be needed. Good tests are simply wasted when used on poor samples. A quick and simple test may suffice for a heterogeneous material represented by one or two samples.

We will suppose that the laboratory has a procedure that appears to be satisfactory. What supporting evidence does the committee have a right to expect from the initiating laboratory? It is useless to exhibit an array of corresponding results obtained on aliquots of a sample or by one operator doing his very best to "hold everything constant." What is needed is positive evidence that the results check acceptably when deliberate variations are made in the test conditions. These variations should be of the size likely to be encountered when several laboratories are presumably following the procedure.

As an example, suppose that the samples have to be placed in an environment of specified humidity and temperature for a certain period of time. The initiating laboratory may subject a dozen samples to this conditioning and obtain excellent checks. A dozen samples sent one each to twelve laboratories yield twelve results with considerable scatter. The explanation is simple. Let the required temperature be 80 C, the relative humidity 60 per cent, and the time 1 hr. Suppose the initiating laboratory for its test sets 78 C, 55 per cent humidity and takes them out after 56 min. Of course, the twelve results still check each other nicely—and this proves nothing at all except that the sampling is adequate. Twelve laboratories will set up various temperatures and humidities, all nominally as specified, and be variously inexact about the time, and this may explain the scatter of the results.

The initiating laboratory has the responsibility to vary the test conditions from the nominal specified values to find out what happens. The initiating research often makes use of better equipment and controls than are available routinely. The initiating laboratory should be able to set 80 C, or 78 C, or some other nearby value and hold it there. Likewise with the other conditions. If the laboratory finds it necessary to set and hold the relevant conditions within very narrow limits in order to achieve good checks this may seriously limit the usefulness of the pro-

NOTE—DISCUSSION OF THIS PAPER IS INVITED, either for publication or for the attention of the author or authors. Address all communications to ASTM Headquarters, 1916 Race St., Philadelphia 3, Pa.

W. J. YOUDEN's academic degrees are in chemical engineering and chemistry. He began to use statistical procedures in 1925 when he was appointed chemist at the Boyce Thompson Institute for Plant Research, Inc. He held this post for 24 years except for the war period when he served as operations analyst with the Air Force. Since 1948 he has been a statistical consultant in the Applied Mathematics Division of the National Bureau of Standards, Washington, D. C. Mr. Youden is the author of more than 100 papers, has written a book (*Statistical Methods for Chemists*), contributed statistical chapters to several other books, and for six years wrote a column "Statistical Design" for *Industrial and Engineering Chemistry*.

TABLE I.—EIGHT COMBINATIONS OF TEST CONDITIONS AND DUPLICATE TEST RESULTS ON 2-IN. CUBES OF CEMENT.

	Combination of Test Conditions							
	1	2	3	4	5	6	7	8
Cement	*A*	*A*	*A*	*A*	*a*	*a*	*a*	*a*
Sand	*B*	*B*	*b*	*b*	*B*	*B*	*b*	*b*
Hours in mold	*C*	*c*	*C*	*c*	*C*	*c*	*C*	*c*
Age at test	*D*	*D*	*d*	*d*	*d*	*d*	*D*	*D*
Initial loading	*E*	*e*	*E*	*e*	*e*	*E*	*e*	*E*
Loading rate	*F*	*f*	*f*	*F*	*F*	*f*	*f*	*F*
Operator	*G*	*g*	*g*	*G*	*g*	*G*	*G*	*g*
Duplicate test results, lb	8200	8680	9100	8620	9620	9540	9160	9320
	8100	8220	9240	8980	9480	9600	9100	9360
Average	8150	8450	9170	8800	9550	9570	9130	9340
Difference	100	460	140	360	140	60	60	40

cedure. Therefore, the initiating laboratory should present evidence to demonstrate that the test procedure results will not be altered by departures from specified values of the test conditions that are likely to be encountered when using routine equipment. To use a round-robin test and hope that all will be well is a misuse of the time of other laboratories. Furthermore, the identification of the particular conditions to which the test results are sensitive is impossible using the round-robin data because naturally all the laboratories report that they followed the specified conditions.

Simple Design for Within-Laboratory Study

The committee should be furnished with actual evidence that the test procedure tolerates departures from specified conditions to the extent that may be expected in practice. Simple and sensitive experimental designs are available for the use of a laboratory undertaking to supply this sort of evidence. Clearly, the selection of the conditions to be explored will depend on the material and on the test procedure. This experimental design is so economical and efficient that the laboratory can include conditions which it might ordinarily feel could safely be assumed not to be a source of trouble. The larger the number of conditions explored the more convincing will be the evidence submitted to the committee.

The principle of the experimental design will be developed in a simple example involving a test of just three conditions. Let the specified values for the conditions be *A*, *B*, and *C* and the alternative values, slightly different from the specified values, be *a*, *b*, and *c*. The standard experimental procedure would be to conduct four trials as follows:

Trial	Condition	Observed Result
No. 1	*A B C*	*t*
No. 2	*a B C*	*u*
No. 3	*A b C*	*v*
No. 4	*A B c*	*w*

[1] The boldface numbers in parentheses refer to the list of references appended to this paper.

The thought here is that, by varying one condition at a time, the effect of changing a condition will be directly revealed. This is true, but there is a more efficient way to conduct the investigation. The four trials listed below are more efficient in detecting possible effects of changing a condition.

Trial	Condition	Observed Result
No. 1	*A B C*	*t*
No. 5	*a b C*	*x*
No. 6	*a B c*	*y*
No. 7	*A b c*	*z*

Notice that *two* conditions have been changed each time from the initial set of conditions *A*, *B*, and *C*. The effect of changing condition *A* to *a* is given by taking the difference between the averages of two results.

$$\text{Average result with } A \ldots \frac{t+z}{2}$$

$$\text{Average result with } a \ldots \frac{x+y}{2}$$

The two trials with condition *A* involve *B*, *b*, *C*, and *c*. This is also true for the two trials at condition *a*. Thus, the effects associated with *B*, *b*, *C*, and *c* are present in both averages, although in the combinations *BC* and *bc* for *A*, and *bC* and *Bc* for *a*. The effect of changing from *B* to *b* is taken to be independent of the value set for condition *C*. The justification rests on the expectation that the changes, *A* to *a*, *B* to *b*, and *C* to *c*, have been made quite small, and therefore the changes are not expected to have an appreciable effect on the test result if the test procedure is acceptable for routine work where such small changes in the conditions are likely to be encountered. If the effect on the test result of changing any capital condition to its lower-case counterpart is substantial, the test procedure is in trouble anyway. If we were trying to establish how a test result changes when some test condition, say temperature, is varied over a very large range, then the interdependence with other conditions would be very important and the proposed design would not be suitable.

The fact is that a good test procedure must not be too sensitive to inadvertent small departures from the specified test conditions. Presumably, there will be small consequences of such departures, consequences not much larger than the experimental error and therefore difficult to detect. The use of the averages, instead of the difference between single tests, gives the investigator a better chance to pick up the effects of departures from the specified conditions. Furthermore, it is altogether reasonable to use as a means of estimating the "error" of the test procedure the variation among the four results *t*, *x*, *y*, and *z*. Not only is it reasonable but more realistic, because surely the performance of the test procedure is given by results of setting up the conditions several times and not from several specimens all exposed to exactly the same conditioning, whatever it happened to be.

Indeed, two or more specimens should be included in each of the four trials and the error within trials (pooled for all trials) compared with that found *between* trials. The committee can, as a minimum, expect to be furnished the between-trial figure for the error, because the results from different laboratories will not be any better than this error and more than likely will be worse.

Illustrative Study of a Test Procedure

Twenty-five years ago Yates (3)[1] proposed such "weighing designs" but considered them of merely academic interest because the agricultural investigations that he was familiar with generally involved large effects on the crop yields. The following example using actual data is based on the design he proposed for seven experimental factors or conditions (8). This investigation concerned the study of seven conditions that might influence the compressive strength of 2-in. cubes of portland-cement mortar. The conditions were: choice of cements, choice of sand, choice of hours in mold, choice of age at test, choice of initial loading versus no initial loading, choice of fast or slow loading rate, and choice of operators. These seven conditions were assigned values and identifying letters as follows:

Cement	*A* or *a*
Sand	*B* or *b*
Hours in mold (16 or 24)	*C* or *c*
Age at test (65 or 72 hr)	*D* or *d*
Initial loading (yes or no)	*E* or *e*
Loading rate (fast or slow)	*F* or *f*
Operator (Joe or Jack)	*G* or *g*

Eight combinations of test conditions are shown in Table I together with the breaking strengths of the duplicate specimens tested for each of the eight combinations. This table shows that combinations 2 through 8 all differ from the

standard combination 1 in that four conditions are changed simultaneously. The changes are made in such a way that the four capital-*A* combinations contain two capital and two lower-case letters for each of the other six letters. The four lower-case-*a* combinations also contain two capital and two lower-case combinations of the other six letters. Thus, the effects of these other letters are balanced off against one another when the average of the four combinations containing *A* is compared with the average of the four combinations containing *a*. This state of affairs holds no matter which letter is selected to determine the division into two groups.

For example, to compare the strengths of the specimens tested after 65 hr with the specimens tested after 72 hr tabulate the four results for *D* and the four results for *d*.

	D (65 hr)	*d* (72 hr)
	8 150	9 170
	8 450	8 800
	9 130	9 550
	9 340	9 570
Total	35 070	37 090
Average	8 768	9 272
Difference	504	

It is not surprising to find that specimens tested after 72 hr are stronger than specimens tested after 65 hr.

The inclusion of variables that would be expected to have little or no effect will provide direct assurance that differences on the order of 500 are meaningful. Thus, the hours in the mold were either 16 or 24.

	C (16 hr)	*c* (24 hr)
	8 150	8 450
	9 170	8 800
	9 550	9 570
	9 130	9 340
Total	36 000	36 160
Average	9 000	9 040
Difference	40	

In spite of the use of different cements, sands, and testing ages all of which influence the strength, the two averages representing different times in the mold show excellent agreement. The above results and those for the other five factors are shown in Table II.

This example is a severe test of this method of studying a test procedure. The 10 per cent change in age is greatly in excess of any expected departure from the test conditions. Indeed, two cements might gain strength at different rates. This complication would usually not be present if only one cement were used. The two cements and the excessively different ages were used to make sure that the example would have at least two sizable differences among the comparisons. A satisfactory procedure should give only insignificant differences when modest and reasonable variations are permitted in the test conditions.

TABLE II.—COMPARISON OF RESULTS OBTAINED WITH CHANGED TEST CONDITIONS. BREAKING STRENGTHS, LB.

Condition Changed	Average for Capital Letters	Average for Lower-Case Letters	Difference Between Averages
Cement	8642	9398	756
Sand	8930	9110	180
Hours in mold	9000	9040	40
Age at test	8768	9272	504
Initial loading	9058	8982	76
Loading rate	8960	9080	120
Operator	8912	9128	216

The problem confronting the investigator is the evaluation of the differences listed in the last column of Table II. The differences between the duplicate specimens (Table I) do provide a basis for judgment for all the conditions except cement and sand. The reason for these two exceptions is that the duplicate test specimens always came from the same batch. Comparisons between hours in the mold, ages at test, initial and no initial loading, between loading rates, and between operators use specimens from the same batch. On the other hand, cements (or sands) cannot be compared without making different batches. Consequently, the reproducibility of the batches is involved, and this may make the comparison of sands and cements subject to a larger error than the duplicate specimen error.

The examination of an experimental situation to identify the possible sources of error applicable to any particular comparison is an often overlooked step in the examination of experimental results. If the effect of changing sands is to be justly evaluated, then a number of repeat batches with each sand should be made. The difference found between batches made with different sands can only be judged by the difference found between batches made using the same sand. In the case at hand the change in strength due to changing the sand is small, indicating that both the change in sand and the difference between batches had small effects. The large effect of changing the cement can therefore be judged to arise mainly from the change in cement itself rather than the nonreproducibility of batches.

The eight pairs of duplicate specimens provide an estimate, *s*, of the standard deviation of a result on a single test specimen. This estimate is based on only 8 degrees of freedom. Triplicate specimens would provide 16 degrees of freedom and, in general, 16 or more degrees of freedom are advisable. The differences, 100, 460, . . . , 40 are squared and divided by $2 \times 8 = 16$, that is, twice the number of pairs.

$$s = \sqrt{\Sigma d^2/16} = \sqrt{399{,}200/16} = 158 \text{ lb}$$

The estimated standard deviation of a *difference* between two averages, when each average is based on *n* results, is $s\sqrt{2/n}$. The averages listed in Table II are based on eight specimens because duplicate cubes were averaged to get the result for each combination. The standard deviation for the last five differences listed in the last column of Table II is $158\sqrt{2/8}$, or 79 lb. The multiple, *t*, of this standard deviation that is taken to give a difference not likely to be exceeded by chance depends on the level of probability selected by the investigator and also on the number of degrees of freedom available for estimating the standard deviation. At the 1 per cent level, with 8 degrees of freedom for the estimate, the value for *t* is 3.36. Consequently, differences of the order of 3.36×79, or about 265 lb, suggest that changing the condition did have an effect. Changing the mold time, operators, the initial loading, and the loading rate all produced smaller differences. With more specimens and firmer averages these differences might be established as something other than fortuitous. The age at test is clearly important at least for this early age. Assuming a linear increase in strength over the interval between the two ages of test (65 and 72 hr), then the 504-lb increase in 7 hr suggests that 1 hr would make a difference of about 70 lb in strength. Clearly the specific time must be adhered to.

There are other factors that might have been studied for their effect on the strength of cubes. For instance, during the time "hours in mold" the specimens in molds are stored in a moist cabinet maintained at 73.4 ± 3 F and not less than 90 per cent humidity. Also the mixing must be done in a temperature between 68 and 81.5 F and a humidity of not less than 50 per cent. The temperature of the mixing slab, dry materials, mold, and mixing bowl are also supposed to be between the latter limits, and the temperature of the mixing water is specified the same as that of the moist cabinet. After the specimens are removed from the molds, they are stored under water, also maintained at 73.4 ± 3 F. It is specified that

the water in the storage tank should be kept "clean by frequent changing." Some people feel that too frequent changing leeches the specimens and changes the strength.

Additional Experimental Designs

Designs to study fewer than seven conditions are easily constructed from the schedule shown in Table I. If only five conditions are to be studied, simply note the identifying labels for *F* and *f*, and *G* and *g*, but make no condition changes for these symbols. The reason for retaining the symbols is that the separation of the eight results into two groups should still be made for each of these letters. The two averages for *F* and *f* ought to agree, within the experimental error, because no change in condition was connected with the grouping. The averages for *G* and *g* should also agree. This provides a desirable check on the experimental error as revealed by the duplicate (or more) results obtained for each of the eight combinations. Incidentally, interchanging the capital and lower-case letters in Table I gives a quite different selection of eight combinations that possesses all the properties of the set shown in Table I. The conditions retain their assigned letters. Should this second set also be tried, a second set of differences which estimate the effects associated with the changed conditions becomes available. This would provide additional confirmation of any effects indicated by the first eight combinations.

Any number up to eleven conditions, *A* through *K*, can be studied by forming twelve combinations using the schedule shown in Table III. This schedule of combinations is from a paper by Plackett and Burman (2) that also lists schedules for larger numbers of combinations.

Error of a Test Procedure

The sponsor of a test procedure should make every possible effort to simulate, in his own laboratory, the sources of error, that is, the changes in conditions that will be encountered in different laboratories. In the cement example, different laboratories unavoidably use different batches, and it is, therefore, the reproducibility of the batch that is involved and not merely the agreement shown by duplicate specimens from the same batch. Of course, in studying within one laboratory the effect of changing certain conditons, such as operator or loading rate, there is a real advantage in making comparisons between specimens from the same batch. This was true in the above example. If the effects are negligible when judged in terms of duplicates from the same batch, they will certainly not matter when the error of different batches is also involved, as it is in comparisons between laboratories.

There have been many attempts to define precision and accuracy and the newer terms, repeatability and reproducibility. The case history just discussed shows how the appropriate error term depends upon the actual situation. It is an oversimplification to talk about within laboratory error and between-laboratory error. Men have had little success in framing definitions acceptable to a majority within a committee and even less success in framing definitions acceptable to a majority of ASTM committees. Perhaps we should worry less about defining these terms and concentrate more on devising some set of operations that will readily reveal the vicissitudes to which a test procedure will be exposed. In addition, there is needed a plain statement of the variation exhibited by the test results—say, the standard deviation—when test conditions are purposely varied. At best such an estimate of the performance of the test procedure is likely to be somewhat optimistic, because the initiating laboratory may have neglected to vary certain conditions or varied some of them by too small amounts. It does seem as though it might be relatively easy to devise an acceptable routine for getting data on a proposed test procedure that bears some relation to the real world of testing. No difficulty stands in the way of selecting a statistical technique that will provide a concise representation of the variation among the results. It should be easier for committes to agree on the operations, both laboratory and statistical, than to agree on the meanings of the words, both old and new, that have served as abstract labels.

The two sands and the two cements in the cement test forced the preparation of four batches, and these were used for the eight combinations. Given a batch for each of the eight combinations the eight batches would give more information on variation arising from batch-to-batch differences. (The cement contrast normally would be used for some test condition.) The laboratory sponsoring the test procedure implies that the test results are not unduly altered by small, unavoidable departures from the specified test conditions. The laboratory should explore such reasonable and inevitable departures. If the sponsoring laboratory believes that it has a satisfactory test procedure, it should be willing to list the eight averages (they may be single results) for the eight combinations and claim no better performance than the standard deviation calculated from these eight results associated with the eight combinations.

If this standard deviation is unacceptably large, then the comparisons listed in Table II should indicate the conditions chiefly responsible. Improved means for setting this condition at its standard value must be devised, or at the very least, the procedure must contain a warning that special, not routine, care is necessary on this condition. All this seems to be a minimum amount of information that should accompany a test procedure under consideration for interlaboratory test. The sponsoring laboratory may have all the fun it wants within its own walls by using nested factorials, components of variance, or anything else that the workers believe will help in the fashioning of a test procedure. At some time the chosen procedure should undergo the sort of mutilation that results from the departures from the specified procedure that occur in other laboratories. The extent of these departures must be based upon expert knowledge of the available equipment and how it is used in routine practice. If the procedure passes this test, it is ready to undergo an interlaboratory test. The interlaboratory test should be a *confirmation* of the claims made for the procedure. The disappointing results so often obtained in round robins are disappointing only in terms of false hopes that were based on unrealistic claims made for the procedure by the sponsoring laboratory.

> "Reproducibility is desirable, but it should not be forgotten that it may be achieved just as easily by insensitivity as by an increase in precision."
>
> Example: "All men are two meters tall give or take a meter."
>
> ANON.

TABLE III.—SCHEDULE FOR TWELVE COMBINATIONS OF ANY NUMBER UP TO ELEVEN CONDITIONS.

1	2	3	4	5	6	7	8	9	10	11	12
A	*A*	*a*	*A*	*A*	*A*	*a*	*a*	*a*	*A*	*a*	*a*
B	*b*	*B*	*B*	*B*	*b*	*b*	*b*	*B*	*b*	*b*	*B*
C	*C*	*C*	*C*	*c*	*c*	*c*	*C*	*c*	*c*	*C*	*c*
D	*D*	*D*	*d*	*d*	*d*	*D*	*d*	*d*	*D*	*d*	*D*
E	*E*	*e*	*e*	*e*	*E*	*e*	*e*	*E*	*e*	*E*	*E*
F	*f*	*f*	*f*	*F*	*f*	*f*	*F*	*f*	*F*	*F*	*F*
G	*g*	*g*	*G*	*g*	*g*	*G*	*g*	*G*	*G*	*G*	*g*
H	*h*	*H*	*h*	*h*	*H*	*h*	*H*	*H*	*H*	*h*	*h*
I	*I*	*i*	*i*	*I*	*i*	*I*	*I*	*I*	*i*	*i*	*i*
J	*j*	*j*	*J*	*j*	*J*	*J*	*J*	*j*	*j*	*j*	*J*
K	*k*	*K*	*k*	*K*	*K*	*K*	*k*	*k*	*k*	*K*	*k*

The Interlaboratory Test

A vast amount of testing time has been wasted upon over-elaborate interlaboratory test programs on procedures whose shortcomings would have been revealed by a modest round robin. This section will present briefly a compact program that will quickly assay the claims made for the test procedure. Should the procedure survive this phase, a more searching and necessarily more elaborate interlaboratory program may be undertaken if considered necessary by the committee.

The proposed interlaboratory test requires two samples of about the same nature and value of the property to be tested. These are sent to a dozen or more cooperating laboratories with the request for one test on each sample according to the test procedure. It is recommended that a second pair of samples quite different in value of the property from the first pair also be circulated. Even then each participating laboratory is asked for only four test results.

The very modest assignment per laboratory should make it feasible to increase the number of participating laboratories and improve the basis for judging the performance obtained by different laboratories.

The elimination of duplicates, the restricted number of materials, and the avoidance of the usual falderal of operators, days, etc., introduces an immense simplification. The committee would be very pleased if the reports from a dozen or more laboratories showed excellent agreement with perhaps one or two exceptions. Automatically the results have sampled equipment, days, operators, etc. If the results show acceptable agreement, that is good. If the agreement among the results is not acceptable the method is unsatisfactory and the claims of the sponsoring laboratory have not been confirmed. In other words, the initiating laboratory has not fully explored the possible sources of variation in the place where such effects are most easy to uncover, namely, in its own laboratory.

Much can be learned from a graph prepared using the pairs of results reported for two closely similar samples. Call these samples X and Y. Lay off x and y axes on a graph using a scale so that the lowest and highest values can be plotted for each sample. The same unit of scale must be used for both axes. Now take the pair of results reported by laboratory A for samples X and Y, and using these two results as coordinates plot a point marked A on the graph paper. Do this for each laboratory until a pattern of points appears on the paper, one point for each laboratory.

Plot another point using the average values for X and Y as coordinates. Draw through this point horizontal and vertical lines dividing the area into four quadrants. If chance errors alone were present in the results, the combinations plus-plus, plus-minus, minus-plus, and minus-minus of random errors would all have the same chance of occurring, and the points would be distributed in a circular pattern around the center with approximately equal numbers of points in each quadrant. The radius of this circular pattern is related to the over-all standard deviation of the test results, sometimes designated as the "reproducibility" of the test.

Usually, however, the points do not form a circle, but a majority of them, and not infrequently nearly all of them, fall in the upper right and lower left quadrants, and more or less close to a line through the center making a 45-deg angle with the X axis. The excess of the plus-plus and minus-minus combinations reflects the presence of some departure from the prescribed conditions for performing the test that carries the same effect over into both results. If this effect is large enough when superimposed upon the small random errors of duplicates, the two results will both be high (or both low) with respect to the grand averages for the two samples.

The points may form a broad oval cluster with only a small excess of points in the plus-plus and minus-minus quadrants. If there are one or two points definitely apart from the cluster and near the 45-deg line, the conclusion may be drawn that these outlying laboratories have failed in some important respect to achieve the specified test conditions. The points are sometimes spread along the 45-deg line in a long narrow oval indicating that nearly all the laboratories were departing from the prescribed conditions. This may come about because the prescribed conditions have not been clearly set forth in the procedure, particularly in the matter of of how closely the standard conditions must be achieved. The procedure may be so vulnerable to even the smallest departures for some of the conditions as to make it impractical for routine use.

The second pair of samples is used for a second graph. Comparison of the two graphs will reveal whether the performance of the procedure changes markedly with the value of the property. If the same laboratory occupies the same extreme position along the 45-deg line on both graphs, this confirms the departure from the specified procedure. A laboratory with points well removed from the clusters but not near the 45-deg line is presumably not even maintaining control of some important conditions. Examples of these two sample graphs are shown in Fig. 1. The reader may make his own interpretation based on the two preceding paragraphs. Detailed accounts of this technique of presenting the results of interlaboratory tests have been published (4–7) and applied to a wide variety of tests.

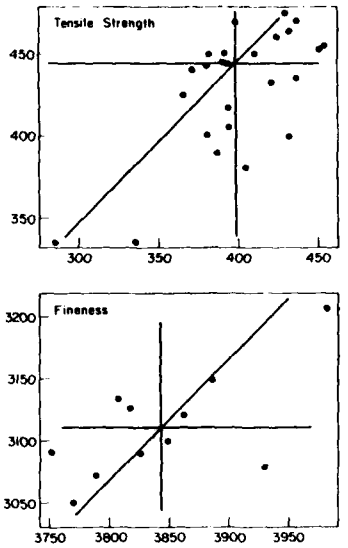


Fig. 1.— Each graph shows two materials tested by several laboratories.

(*Top*) Results for tension tests (psi). (*Bottom*) Results of tests for fineness of cement (sq cm per g). The pair of results reported by a laboratory are used to plot a point. The x axis is used for the result reported on one material, the y axis for the result reported for the other material. In each case one or two laboratories are clearly apart from the main cluster of points.

Evaluating the Quality of the Test Procedure

The scatter of the points plotted in the two sample diagrams directs attention to a responsibility all too often shirked by those entrusted with the evaluation of test procedures. The diagrams in Fig. 1 and other diagrams in the cited papers have one or more points clearly apart from the main cluster. What disposition is to be made of the results that are responsible for these outlying points? If the between-laboratory error is calculated using the data from all the laboratories the error is considerably inflated by the retention of the results associated with these points. One answer to the above question is to use all the data to establish the performance of the method on the ground, that, among the laboratories not participating, there may be a few more like the one or two responsible for the outlying points appearing in the diagram. This would appear to put the emphasis on the performance of the laboratories rather than on the inherent quality of the procedure when properly used. The other answer will require directing the attention of all concerned to those laboratories whose

pronounced individuality sets them apart from the overwhelming majority. These laboratories would have the alternatives of justifying their values, or discovering the causes of their troubles and removing them, or of being quietly omitted from the group used to evaluate the procedure. No amount of discussion about accuracy, however prolonged, and no statistical techniques, however complicated, can be substituted for a straightforward facing up to the problem of these outlying laboratories.

The problem of outlying results confronts all those concerned with the improvement of test procedures and all who use these procedures. The statistician can assist the engineers after they have settled in their minds what it is they want. If the decision is made to retain all the data, except clearly bizarre results, the setting of confidence limits may be relatively meaningless. To ask the statistician to make some prediction about a new laboratory is to invite the reply "Is it a good laboratory or a bad one?" And we are right back where we have always been. If the decision is made to set aside some of the results, should it appear necessary, then the statistician can be of considerable assistance in respect to the rules for eliminating results. Confidence limits for points in the main cluster can be set with some assurance that they apply to laboratories of the same competence as those in the main cluster.

It is interesting that in other activities, such as passing a college examination, a standard is set that a large majority of the students can meet successfully. No one is disturbed that some fail for lack of application or equipment. In a very real sense the situation is closely parallel to the performance of the laboratories with a test procedure. Assign a large standard deviation and all the laboratories get in. But an examination that everybody can pass does not do justice to the course nor does it reveal its actual merit. The committees must come to grips with this problem—no one else will.

Acknowledgments:

The author is greatly indebted to Howard T. Arni of the Inorganic Building Materials Section of the National Bureau of Standards who made numerous helpful suggestions. The data in Table I were obtained through the cooperation of Mr. D. N. Evans of the same section who arranged to have the tests made to illustrate the application of this statistical design.

References

(1) J. R. Crandall and R. L. Blaine, "Statistical Evaluation of Interlaboratory Cement Tests," *Proceedings*, Am. Soc. Testing Mats., Vol. 59, p. 1129 (1959).

(2) R. L. Plackett and J. P. Burman, "The Design of Optimum Multifactorial Experiments," *Biometrika*, Vol. 33, p. 305 (1946).

(3) F. Yates, "Complex Experiments," Royal Statistical Society, *Supplement*, Vol. 2, p. 181 (1935).

(4) W. J. Youden, "Statistical Aspects of the Cement Testing Program," *Proceedings*, Am. Soc. Testing Mats., Vol. 59, p. 1120 (1959).

(5) W. J. Youden, "Statistical Design," *Industrial and Engineering Chemistry*, Vol. 50, Aug., 1958, p. 63A; Oct., 1958, p. 91A; Dec., 1958, p. 77A.

(6) W. J. Youden, "Evaluation of Chemical Analyses on Two Rocks," *Technometrics*, Vol. 1, p. 409 (1959).

(7) W. J. Youden, "Graphical Diagnosis of Interlaboratory Test Results," *Industrial Quality Control*, Vol. XV, p. 24, May, 1959.

(8) W. J. Youden, "Statistical Design," *Industrial and Engineering Chemistry*, Vol. 51, p. 79A, Oct., 1959.

Ranking Laboratories by Round-Robin Tests

By W. J. YOUDEN

This paper presents a method for scoring laboratories participating in round-robin tests. For each material the laboratory with the highest numerical result is given the rank of one, the laboratory with the next highest result is given the rank of two, and so on until the lowest result is given the lowest rank (L). A laboratory is scored by summing its ranks for all the materials. The paper includes a new statistical table that gives lower and upper limits for scores that correspond to 5 per cent probability. Because systematic errors produce extreme scores, the table should be useful in singling out laboratories with pronounced systematic errors.

ROUND ROBINS are undertaken for a variety of motives: (1) to accumulate data that may be used to determine the precision and accuracy of a new or modified test procedure, (2) to recheck an established procedure to ascertain whether there has been a deterioration in the accuracy arising from departures from the prescribed routine, (3) to test the applicability of an established procedure to new materials, and (4) to maintain a periodic check on the performance of a group of laboratories.

The questions to be answered by a round robin depend on the information already in hand. The procedure may be a new test worked out in one laboratory. Usually this laboratory has data that should provide a fair estimate of the agreement that can be obtained between measurements made under the same conditions. This laboratory has the exacting task of preparing an adequate description of the apparatus, environment, and technique for making the measurements. Failure to include relevant items can be quite disastrous, particularly if the items are fixed and unchanging in the originating laboratory. There is the risk of wasting much effort if a full-scale round robin reveals a diversity among the results that can only be explained by shortcomings and ambiguities in the instructions for performing the test. The originating laboratory can and should check the homogeneity of the samples so that any unsatisfactory results cannot be ascribed to sample difficulties.

The only positive way to check the adequacy of the instructions is to ask other laboratories to try the procedure. Single results on two rather similar materials by seven or more laboratories should catch any major shortcomings in the instructions. If the between-laboratory error is several times as large as the precision established by the originating laboratory, some of the laboratories are probably unintentionally deviating from the routine followed in the originating laboratory. This conclusion may be further checked by listing the difference between the results for the two materials as reported by each laboratory. If the seven differences show much better agreement among themselves than do the seven values reported for the first material, some, or all, of the laboratories probably have individual interpretations of the procedure. Usually such individual interpretations have the same effect on the results for both the materials. Since systematic effects drop out when the differences are taken, the differences will show better agreement than the actual values. This examination of the data may show that it is necessary to rewrite the instructions before collecting a large mass of data.[1]

The originating laboratory also should establish the range of materials for which the procedure gives satisfactory results. Once it has been shown that the procedure is stated properly, a more limited number of laboratories, even as few as two, may test perhaps a dozen or more materials for which the property under test varies widely. Now the results obtained by one laboratory should be subtracted from the corresponding results obtained by the other laboratory. These differences, if almost all of one sign, indicate that one laboratory is biased relative to the other. Each difference (taking account of sign) may be plotted against the corresponding average of the two results. One should look for some pattern, such as differences that tend to increase in size with increases in the magnitude of the average. If the procedure passes this check on the claims of the originating laboratory, a more comprehensive program may be undertaken.

Sometimes the purpose of a round robin is to determine whether it is necessary to maintain a stock of standard or reference samples so that laboratories may check their equipment and technique. The results with reference samples form the basis for adjustments to the equipment or for making arbitrary corrections to routine test results. Procedures that require this prop are usually troublesome and expensive.

Interpretation of Data

Comprehensive round robins involving a considerable number of laboratories often yield collections of data that pose problems in evaluation. If trouble turns up with carefully selected laboratories, the procedure or the adequacy of its description is already suspect. If the procedure still shows promise after these preliminaries, it remains to be shown

NOTE—DISCUSSION OF THIS PAPER IS INVITED, either for publication or for the attention of the author or authors. Address all communications to ASTM Headquarters, 1916 Race St., Philadelphia 3, Pa.

[1] W. J. Youden, "Experimental Design and ASTM Committees," *Materials Research & Standards*, Vol. 1, No. 11, Nov., 1961, pp. 862-867.

W. J. YOUDEN is a chemical engineer by training. After 20 years of laboratory research at the Boyce Thompson Institute for Plant Research, Inc., he joined the staff of the National Bureau of Standards. His special interest lies in the design of experiments. His latest book, *Experimentation and Measurement*, is the second in a series published by the National Science Teachers Assn. Mr. Youden has been active on several ASTM committees. His experiences with these committees has made him aware of the need for the statistical technique described in this paper.

that good results can be obtained with a random selection of laboratories.

A round robin that takes in a cross-section of typical laboratories goes beyond an evaluation of the procedure. The data that will be collected reflect the merits of the procedure and also reflect the performance of the participating laboratories. Poor results may be caused by deficiencies in the procedure or failures to follow the procedure faithfully. Judgment of the procedure will be made on the data remaining after deleting absurd results. It will be shown that one deviant laboratory can easily account for a considerable fraction of the sum of the squared deviations used in evaluating the error.

Rejection of Results

There has long been needed some guide or aid to the judgment in those difficult situations that accompany the rejection of results submitted by a laboratory. What is needed is some understandable criterion that is convincing even to the laboratory concerned. For example, suppose a round robin involves nine laboratories testing seven materials. All the laboratories measure the same property on all the materials. Imagine that one of the laboratories turns in the highest (or lowest) result for every one of the seven materials. This event cannot be ascribed to chance. If the ace and all diamonds up to and including the nine spot are removed from a deck of cards and shuffled, the laboratory concerned may be challenged to pick the ace when the nine cards are spread face down. All the laboratory has to do is succeed in this effort seven times in succession, the cards being reshuffled each time. It is not necessary to mention the odds against achieving this performance. Even if the laboratory representative succeeded two or three times in succession, many would suspect that the cards were marked on their backs. That is, everyone would soon conclude that there was something to be explained. And that is just the point. The laboratory should explain why it gets such extreme results. Short consideration can be given the suggestion that these extreme results may be correct and the other eight laboratories share a common error. Conceivably that may happen but why go against the majority? It seems only reasonable to put the burden of proof on the single laboratory rather than on the other eight.

A general criterion for rejection of results could consist of assigning the ranks one to nine to the nine results reported by the nine laboratories on the first of seven materials. If multiple tests have been made on the same material, the average of the results is used to represent the laboratory. The rank of one goes to the laboratory with the highest result, the rank of nine to the laboratory with the lowest result. If a tie exists, say two laboratories are tied for fifth place, assign the rank of 5.5 to the tied laboratories. If three are tied for fourth place, assign the middle rank of five to all three. This maintains the total of the ranks at 45 for each material. The average rank is 45/9 or 5.

When the laboratories have ranks assigned for all seven materials, a score is given each laboratory by adding up its ranks. A score of seven is the minimum possible (highest every time), and a score of 63 is the maximum possible (meaning that the laboratory reported the lowest result on every material). The average score is 7×5, or 35, just midway between the minimum and maximum. If only random errors were involved, the rank a laboratory got on each material would be simply a matter of chance. To get an idea of the scores that turn up, shuffle nine cards (ace through nine of diamonds) to get them in random order, and then write the numbers opposite the letters A to I that identify the nine laboratories. Repeat this process until seven ranks have been entered against each letter. Sum the ranks and observe the scores. The outcome of such a simulated round robin is shown in Table I.

This game was tried 1000 times with the aid of a computer. Examination of all 9000 scores shows that there were 22 scores of 16 or less and

TABLE I.—RANDOM ARRANGEMENTS OF NINE CARDS.

Laboratory	Trial No. 1	2	3	4	5	6	7	Score
A	4	2	6	5	9	1	4	31
B	6	8	1	3	1	9	6	34
C	3	6	4	7	8	7	5	40
D	1	4	8	2	4	3	7	29
E	7	9	9	6	6	6	1	44
F	9	5	3	1	3	4	3	28
G	2	1	5	8	5	8	2	31
H	5	3	7	9	7	2	8	41
I	8	7	2	4	2	5	9	37
							Average	35
Sum of squares of differences from average score								264

TABLE II.—SCORES FROM 20 SIMULATED ROUND ROBINS WITH 9 LABORATORIES AND 7 MATERIALS.

29	28	32	49	36	40	38	44	38	21	39	24	35	29	39	26	38	34	43	29
37	40	39	41	30	37	26	35	33	30	40	54	22	38	31	26	32	35	35	44
37	31	34	45	55	37	29	35	29	30	28	36	39	40	40	30	37	32	29	45
35	32	37	27	28	42	51	32	36	37	38	33	26	41	33	34	31	36	42	37
29	37	36	25	33	31	38	27	43	38	43	35	31	24	36	39	23	29	48	25
42	54	37	31	35	31	31	46	29	41	23	34	32	32	36	44	43	41	26	38
23	31	26	27	36	33	34	38	37	34	38	36	44	41	26	39	47	41	25	34
45	27	43	36	28	20	28	29	32	45	34	23	48	26	27	28	33	29	26	28
38	35	31	34	34	44	40	29	38	39	32	40	38	44	47	49	31	38	41	35

TABLE III.—APPROXIMATE 5 PER CENT PROBABILITY LIMITS FOR RANKING SCORES.

Number of Laboratories Participating	Number of Materials 3	4	5	6	7	8	9	10	11	12	13	14	15
3	4	5	7	8	10	12	13	15	17	19	20	22	
	12	15	17	20	22	24	27	29	31	33	36	38	
4	4	6	8	10	12	14	16	18	20	22	24	26	
	16	19	22	25	28	31	34	37	40	43	46	49	
5	5	7	9	11	13	16	18	21	23	26	28	31	
	19	23	27	31	35	38	42	45	49	52	56	59	
6	5	7	10	12	15	18	21	23	26	29	32	35	
	23	28	32	37	41	45	49	54	58	62	66	70	
7	5	8	11	14	17	20	23	26	29	32	36	39	
	27	32	37	42	47	52	57	62	67	72	76	81	
8	6	9	12	15	18	22	25	29	32	36	39	43	
	30	36	42	48	54	59	65	70	76	81	87	92	
9	6	9	13	16	20	24	27	31	35	39	43	47	
	34	41	47	54	60	66	73	79	85	91	97	103	
10	7	10	14	17	21	26	30	34	38	43	47	51	
	37	45	52	60	67	73	80	87	94	100	107	114	
11	7	11	15	19	23	27	32	36	41	46	51	55	
	41	49	57	65	73	81	88	96	103	110	117	125	
12	7	11	15	20	24	29	34	39	44	49	54	59	
	45	54	63	71	80	88	96	104	112	120	128	136	
13	8	12	16	21	26	31	36	42	47	52	58	63	
	48	58	68	77	86	95	104	112	121	130	138	147	
14	8	12	17	22	27	33	38	44	50	56	61	67	
	52	63	73	83	93	102	112	121	130	139	149	158	
15	8	13	18	23	29	35	41	47	53	59	65	71	
	56	67	78	89	99	109	119	129	139	149	159	169	

NOTE.—Let L laboratories test each of M materials. Assign ranks 1 to L for each material. Sum the ranks to get the score for each laboratory. The mean score is $M(L + 1)/2$. The entries are lower and upper limits that are included in the approximate 5 per cent critical region.

21 scores of 54 or more for a total of 43 outlying scores. Exact enumeration gave 42.65 as the expected number of such scores. This is just about ½ of 1 per cent of the 9000 scores. There are nine scores per round robin, so the chance of any given round robin having one of these extreme scores is about nine times this ½ of 1 per cent; or about 5 per cent. Although doubling up would reduce the chance of finding a round robin with an extreme score, no such doubling up was found at this probability level.

Table II lists the scores for 20 of the 1000 simulated round robins. They were picked by taking every fiftieth, starting with number 50, of the 1000 computer-simulated round robins. Note the three extreme scores (54, 54, and 55) in three of the round robins, although only one was expected. Since there are only 45 of these 1000 round robins with an extreme score, it was unusual to get three round robins with extreme scores out of 20 selected in this manner. However, for all other sets of nine scores, the scores stay well within the range from 16 to 54. The scores cluster fairly closely around the average score of 35.

Table III lists the corresponding 5 per cent probability limits for various combinations of number of laboratories and number of materials. All of the results were obtained by direct enumeration of the actual probabilities of getting the indicated lower limit or less and the indicated upper limit or more. Because the scores go by units, it is not possible to have them correspond to the exact 5 per cent probability level. The tabulated scores in some instances correspond to a probability somewhat more than 5 per cent and in other cases to a smaller than 5 per cent limit. The probability refers to the chance of obtaining a round robin with the indicated extreme score.

Combinations for large numbers of both laboratories and materials are not given. The arithmetic became heavy in this region, at least with a desk calculator. More important, there is the question as to how often one is justified in requesting such a large program. If there does seem to be a need to have many laboratories and many materials, the data may be divided on some reasonable basis. Thus many laboratories might be split into two or more groups, say, geographically, or even randomly. Materials could be split into groups on the basis of the magnitude of the property or some other distinctive characteristic.

[3] More extensive tables and further details about this method are contained in the forthcoming paper "A Rank Test for Outliers," by W. A. Thompson, Jr., and T. A. Willke, to be published.

Examples of Scoring Laboratories

The Plant Food Institute sends a monthly sample to a large number of laboratories. The ranks for nine laboratories for seven successive months are shown in Table IV. The choice of nine and seven was made to permit direct comparison with the machine scores shown in Table II. The scores appear very similar to those shown in Table II. Perhaps systematic errors do not persist over several months so there is no unusually low or high score. Even so, laboratory 7 was obtaining low ranks except for the first month, and laboratories 25 and 29 are generally credited with high ranks.

Table V also shows the ranks for another nine laboratories testing seven materials. Actually these laboratories tested 14 materials, but these were split into two groups of seven. The left-hand group of materials are those with low values of the property; the right-hand group includes the materials with high values for the property. The tabulated limits of 16 and 54 are sharply exceeded in both groups. Laboratory 6 comes very close to a clean sweep for rank 9 every time. Laboratory 4 is almost always runner-up to laboratory 6. Laboratories 1 and 2 competed for ranks 1 and 2 in the first group but are in good positions in the second group. There is a pronounced tendency for a laboratory to maintain its position relative to the other laboratories. If this state of affairs cannot be remedied by individual corrective action, the situation may call for the use of reference samples to bring the laboratories into better agreement.

The ranks for 15 laboratories all making determinations of the per cent of indigestible residues on the same seven

TABLE IV.—TOTAL NITROGEN DETERMINATION BY 9 LABORATORIES ON 7 SUCCESSIVE MONTHLY FERTILIZER SAMPLES.

Laboratory	May	June	July	Aug.	Sept.	Oct.	Nov.	Score[a]
No. 7	8	3	2	3	4	3	1	24
No. 8	9	6	1	4	2.5	1.5	3	27
No. 11	6	4	9	6	7	6	4	42
No. 14	1.5	7.5	7	5	5	1.5	2	29.5
No. 16	3	1	6	2	1	7	9	29
No. 25	1.5	9	8	8	6	8	5	45.5
No. 28	7	2	4	9	2.5	5	7	36.5
No. 29	5	5	3	7	8	9	8	45
No. 30	4	7.5	5	1	9	4	6	36.5
							Average	35.0

[a] Critical limits for scores are 16 and 54 (Table III).

TABLE V.—RANKING RESULTS OBTAINED BY 9 LABORATORIES TESTING 14 MATERIALS.[a]

Laboratory	Ranking							Score[b]	Ranking							Score[b]
No. 1	1	1	2	2	2	1	1	10	1	3	3	3.5	7	7	4	28.5
No. 2	2	2	1	1	1	2	2	11	2	2	4	3.5	5	5	3	24.5
No. 3	3	6	3	3	5	7	4	31	3	6	1	6	6	3	7	32
No. 4	5	5	7	8	8	4	8	45	8	8	8	8	8	8	8	56
No. 5	9	3.5	4	4	7	3	3	33.5	4	1	5	1	1	1	2	15
No. 6	8	8	9	9	9	9	9	61	9	9	9	9	9	9	9	63
No. 7	6	9	6	7	4	8	7	47	5	7	7	7	2	4	6	38
No. 8	7	7	8	6	3	5.5	5	41.5	7	4	6	5	4	6	5	37
No. 9	4	3.5	5	5	6	5.5	6	35	6	5	2	2	3	2	1	21
						average		35						average		35

The materials have been grouped on the basis of the magnitude of the property.
Critical limits for scores are 16 and 54 (Table III).

TABLE VI.—RANKING OF 15 COLLABORATIVE RESULTS FOR THE AMOUNT OF INDIGESTIBLE RESIDUES IN 7 PROTEIN MATERIALS.[a]

Laboratory	Materials Analyzed							Score[b]
	SG	MS	PB	BM	DT	MO	MH	
No. 1	8	4	11.5	12	1.5	1	13.5	51.5
No. 2	15	15	1	4	15	15	1	66
No. 3	7	9	15	6	5	10	2	54
No. 4	14	13	14	15	13	14	9	92
No. 6	11.5	8	8.5	3	5	8	3	47
No. 7	6	2.5	6.5	3.5	9.5	11	10	59
No. 8	3	5.5	13	1	7	13	12	54.5
No. 9	11.5	10	11.5	13.5	14	12	5	77.5
No. 10	4.5	7	4.5	8.5	5	5	13.5	48
No. 11	2	2.5	8.5	2	3	6.5	11	35.5
No. 12	4.5	11.5	3	10	1.5	2	15	47.5
No. 13	1	1	2	6	9.5	3	7	29.5
No. 14	9	5.5	4.5	11	8	4	6	48
No. 15	11.5	14	10	8.5	11	6.5	4	65.5
No. 16	11.5	11.5	6.5	6	12	9	8	64.5
							Average	56

[a] Taken from Table I, *Journal*, Assn. of Official Agricultural Chemists, Vol. 42, p. 232, 1959.
[b] Critical limits for scores are 23 and 89 (Table III).

protein materials are shown in Table VI. The scores that are beyond the 5 per cent point are 23 and less, and 89 and more.

The lowest of 15 results reported by the 15 laboratories is given the rank 15. If a laboratory obtained the lowest result on every one of seven materials, its score would be 105. The score for laboratory 4 is 92. The individual ranks are 14, 13, 14, 15, 13, 14, and 9. Evidently this laboratory has a tendency to get lower results than most of the other laboratories. Except for the last material, this laboratory maintains a consistent position in the ranking scale. Evidently this laboratory follows some individual practice in a careful manner. The scores given in Table III should convince laboratory 4 that its string of low values cannot be ascribed to chance. It may be more appropriate to ask laboratory 4 to review its technique rather than to report adversely on the procedure.

The ranks listed in Table VII are interesting because in the left-hand group the collaborators used the method of their own preference in making the determinations. Not all the laboratories participated in both programs, but the same samples were used for both programs. One might have anticipated that some laboratories would maintain their positions relative to the others when the laboratories were invited to use any method they preferred. Laboratory 2 does reach the critical score of 9, and laboratories 4 and 12 approach the other limit of 41. It is more surprising to find that the laboratories show definite individuality when all were presumably following the same tentative procedure. The tentative procedure may be charged with an inflated error that is actually caused by laboratories that are highly individualistic in the way they conduct the test. The lessons to be drawn from a round robin might be immensely helpful to these collaborator laboratories. The author has encountered round-robin data in which the scores were nearly the worst possible: M, $2M$, $3M$, ..., $7M$. The conclusion here is that the description of the procedure does not specify properly some of the test conditions and equipment that influence the test result.

Exceptionally low or high scores support the supposition that the laboratory concerned is doing something uniquely different from the rest. It hardly seems just to the procedure under scrutiny to allow such uniquely different results to inflate the calculated error for the procedure. Extreme scores should prompt the investigator concerned to review the

[illegible] A Comparison of [illegible] for the [illegible] Records of Weather [illegible] 1960.

TABLE VII.—RANKS AND SCORES OF LABORATORIES REPORTING PERCENTAGE OF TOTAL ALKALOIDS AS NICOTINE.[a]

	Collaborator's Choice of Method						Tentative Recommended Method					
	Sample						Sample					
Laboratory	A	B	C	D	E	Score[b]	A	B	C	D	E	Score[c]
No. 1	...	...	...	...	...	...	6	5	7	10	2.5	30.5
No. 2	1	1	2	2	3	9	2	3	1	6	5	17
No. 3	9	9	7	7	8	40	7	8	4	7	10	36
No. 4	7.5	8	9	6	7	37.5	8	6.5	5.5	8	4	32
No. 5	...	...	...	...	...	...	3	1.5	5.5	3.5	1	14.5
No. 6	6	2	4.5	1	1.5	15	4	6.5	9	1	7.5	28
No. 9	4	4	4.5	4	5	21.5	10	10	10	9	9	48
No. 10	2	3	1	3	1.5	10.5	...	...	...	...	...	...
No. 11	5	6	6	8	5	30	1	1.5	2	2	2.5	9
No. 12	7.5	6	8	9	9	39.5	...	...	...	...	...	...
No. 13a	...	...	...	...	...	...	9	9	8	3.5	7.5	37
No. 13b	...	...	...	...	...	...	5	4	3	5	6	23
No. 15	3	6	3	5	5	22						
				Average		25.0				Average		27.5

[a] See *Journal*, Assn. of Official Agricultural Chemists, Vol. 42, p 305, 1959.
[b] Critical scores are 9 and 41.
[c] Critical scores are 10 and 45.

interpretation given the instructions and check possible sources of error.

Discussion

There are fields of work where judges undertake to rank materials in order of merit. Often the judges do not agree among themselves in such subjective tests. A statistical measure of the concordance of the judges has long been in the statistical literature. Fortunately quantitative measurements usually do manage to get the materials in the correct order no matter which laboratory tests the materials. It is not this ranking that has been the object of interest in this paper. Rather the materials may be regarded as ranking the laboratories. If only random errors are operative, the order of the laboratories should not persist from material to material and there should be no concordance whatever.

The goal in the development of a test procedure is to attain an absence of concordance. The ranking scheme is a simple arithmetical device to measure progress toward that goal. If the ranks depend only on chance, the expected sum of squares associated with the scores when L laboratories are ranked M times is $ML(L-1)(L+1)/12$. Denote this sum by S'. Systematic errors spread the scores over a wider range and give a larger sum of squares than S'. Denote the sum of the squared deviations of the observed individual scores from the mean score, $(L+1)M/2$, by S. The ratio S/S' should be distributed approximately as χ^2/f, where f is one less than the number of laboratories.[3] The maximum sum of squares is obtained from the scores $1M$, $2M$, $3M$, ..., LM which indicate perfect concordance. This sum of squares is equal to $M^2(L^3-L)/12$. Dividing this quantity by the expected sum of squares, S', gives M as the maximum value the ratio S/S' can take.

A ratio in the neighborhood of unity is desirable. Ratios less than unity are purely chance occurrences. Because the distribution of the scores is closely approximated by the normal distribution the tabulated values for

TABLE VIII.—PROBABILITY LIMITS FOR THE RATIO OF THE CALCULATED SUM OF SQUARES FOR SCORES TO THE EXPECTED SUM OF SQUARES, $ML(L-1)(L+1)/12$.

Number of Laboratories Participating	Limiting Ratio for		
	5%	1%	0.1%
3	3.00	4.60	6.91
4	2.60	3.78	5.42
5	2.37	3.32	4.62
6	2.21	3.02	4.10
7	2.10	2.80	3.74
8	2.01	2.64	3.47
9	1.94	2.51	3.27
10	1.88	2.41	3.10
11	1.83	2.32	2.96
12	1.79	2.25	2.84
13	1.75	2.18	2.74
14	1.72	2.13	2.66
15	1.69	2.08	2.58
16	1.67	2.04	2.51
17	1.64	2.00	2.45
18	1.62	1.97	2.40
19	1.60	1.93	2.35
20	1.59	1.90	2.31

NOTE.—The above entries were taken from a table of χ^2/f. Exact values for the ratio for certain selected values of L and M are given in Friedman's paper[3]. Friedman's values are almost always slightly smaller than those given above.

TABLE IX.—RATIO OF OBSERVED SUM OF SQUARES S TO EXPECTED SUM OF SQUARES S'.

Table	Laboratories, L	Materials, M	Sum of Squares: Calculated, S	Sum of Squares: Expected, S'	Ratio, S/S'	5 Per Cent Limit
No. I	9	7	264	420	0.63	1.88
No. IV	9	7	515	420	1.22	1.88
No. V	9	7	2181	420	5.20	1.88
No. V	9	7	1995.5	420	4.20	1.88
No. VI	15	7	3030	1960	1.54	1.67
No. VI	14	7	1959.5	1592.5	1.23	1.69
No. VII	9	5	1204	300	4.01	1.88
No. VII	10	5	1254	412.5	3.04	1.83

x^2/f may be used to obtain the approximate upper 5 per cent limit for values of the ratio. Values in excess of the tabulated limits in Table VIII indicate that systematic errors are producing some undesired concordance among the rankings.

The ratio has been calculated for the scores given in the examples in the tables. Thus the random ranks assigned by the playing cards shown in Table I gave scores whose ratio is less than one. The machine-generated scores for the 20 round robins tabulated in Table II gave the following ratios:

First ten: 0.91, 1.29, 0.47, 1.38, 1.26, 1.06, 1.15, 0.85, 0.41, 0.98.

Second ten: 0.78. 1.57, 1.31, 1.03, 0.84, 1.30, 0.98, 0.39, 1.47, 0.91. All of these ratios fall below the upper 5 per cent limit of 1.94 for nine laboratories.

The nitrogen results in Table IV and the data in Table VI gave acceptable ratios (Table IX). Notice that laboratory 4, singled out by the limits given in Table III as having a systematic error, contributed about one third of S. The sum of squares is reduced from 3030 to 1959.5 when this laboratory is dropped. With laboratory 4 included, the probability level for the ratio 1.54 is under 10 per cent.

The data in Tables V and VII yielded large values for all the ratios. Even the smallest of these is close to the tabulated value, 3.10, for a probability level of 0.1 per cent. These large ratios cannot be ascribed to one or two laboratories but are associated with a generally unsatisfactory state of affairs. Perhaps the test procedure is very sensitive to quite minor departures from the specified techniques for performing the test. Or perhaps the instructions are not specific enough. Whatever the reason, most of the laboratories are involved. The remedy here is to give the test procedure a thorough going over by a good laboratory. The effect of intentional deviations from stated conditions for conducting the test should be studied to discover if this accounts for the scatter of the result.[1] Usually any deviation is maintained over long periods, and this would account for a laboratory obtaining about the same rank on all the materials.

Summary

This method of ranking laboratories has advantages besides those of simplicity and ease of calculation. There is no need to be concerned now that the precision may vary from one laboratory to another. Poor precision will tend to invite low or high individual ranks but in equal proportions so there is compensation. Differences in precision are fortunately rather small, or the usual analysis of variance that uses the actual values would run into statistical difficulties. Perhaps the most important advantage of the ranking procedure is that the variance of the scores is known *a priori*. The variance is given by $M\ (L - 1)\ (L + 1)/12$. Indeed, the complete theoretical distribution of the ranks can be obtained if desired. When laboratory averages, obtained from the numerical values, are used to consider the possible rejection of a laboratory, the suspect laboratory average is part of the data and may give such a large estimate for the laboratory variance component that the rejection level is rather generous. With ranks, the rejection levels can be set in advance of seeing any data. Finally, the ranking criterion is intuitively meaningful quite apart from any knowledge of advanced statistical techniques.

Systematic errors that are largely responsible for the disagreements that arise among laboratories probably cannot be completely eliminated. Thus the ranking scores obtained for round robins will tend to cover a wider range than theory predicts. So, too, the ratio S/S' will tend to reach large values. Both the scores and the ratio S/S' provide a convenient measure for gaging improvement. The ratio reflects the general performance of all of the laboratories, whereas the limiting scores focus attention on the laboratories with the extreme scores.

Table III provides an objective criterion for singling out laboratories that have the most pronounced systematic errors. Table VIII provides a quick evaluation of the data as a whole. Once laboratories become convinced that they are deviating from the procedure, the resulting search for the source of the deviation should produce (1) a general improvement in the quality of testing, (2) a better estimate of the inherent quality of the test procedure, and (3) perhaps fewer procedures that appear to require the prop of expensive reference materials.

Acknowledgments:

The author is indebted to William A. Thompson, Jr., for devising a reiterative scheme of computation for Table III. Mary C. Croarkin and Thomas A. Willke checked and extended a small preliminary table obtained by the author by programming the computations on a computer.

The Interlaboratory Evaluation of Testing Methods

By JOHN MANDEL and T. W. LASHOF

Trained manpower and laboratory facilities can be used more effectively if improvements can be made in interlaboratory evaluation of testing methods. There are probably hundreds of these cooperative programs going on all the time under the aegis of the Society's 80 main technical committees. Too often the report of an interlaboratory program indicates the results are not useful because some variable was not adequately controlled or because there was some flaw in planning the program.

Planning interlaboratory test programs has occupied the attention of *most if not all of the technical committees*; in fact, two—D-11 on Rubber and D-13 on Textiles—have prepared recommended practices which have been published by ASTM (D 1421 and D 990, 1958 Book of ASTM Standards, Parts 9 and 10).

Committee E-11 on Quality Control of Materials has the assignment to develop a general recommended practice for interlaboratory testing for use by all the committees. It is accordingly quite interested in the present paper, as indicated by the following statement by one who reviewed the paper for the committee:

"This paper gives a very complete treatment of the problem which almost every ASTM committee is constantly trying to solve . . .(it is) a more comprehensive approach to the problem of designing and interpreting interlaboratory studies than has appeared in the literature up to now. Their (the authors') ideas are complex because the problem they are trying to solve is complex."—ED.

The various sources of variability in test methods are examined, and a new general scheme to account for them is proposed. The assumption is made that systematic differences exist between sets of measurements made by the same observer at different times or on different instruments or by different observers in the same or different laboratories and that these systematic differences are linear functions of the magnitude of the measurements. Hence, the proposed scheme is called "the linear model." The linear model leads to a simple design for round-robin tests but requires a new method of statistical analysis, geared to the practical objectives of a round robin. The design, analysis, and interpretation of a round robin in accordance with the linear model are presented, and the procedure is illustrated in terms of the data obtained in an interlaboratory study of the Bekk smoothness tester for paper. It is believed that this approach will overcome the "frustrations" that are often associated with the interpretation of round-robin test data.

IN THIS paper a new approach is presented for the analysis of interlaboratory studies of test methods. The various sources of variability in test methods are first reexamined and a new general scheme to account for them is proposed. This scheme leads to a simple design for round-robin tests but requires a new method of statistical analysis, geared to the practical objectives of a round robin. The theoretical details are dealt with in a companion paper (1).[1] In the present article, the emphasis is on the application of the new concepts to ASTM committee studies of test methods. The procedure is illustrated in terms of the data obtained in an interlaboratory study of the Bekk smoothness tester for paper.

For much of the discussion in this paper, the consideration of different laboratories is not an absolute requirement. The word "laboratory" is used here to denote a set of measurements obtained under conditions controlled within the set but such that systematic differences may exist from one set to another. For example, different operators within the same laboratory may also show systematic differences. The same may be true for sets of measurements obtained even by the same operator at different times. Since the use of different laboratories, in the usual sense, is likely to result in the greatest number and severity of systematic differences, the practice of conducting interlaboratory round-robin programs for the study of test methods appears entirely justified.

A New Approach: The Linear Model

We will assume that an interlaboratory study of a particular test method has been run in accordance with the schematic diagram shown in Table I; specifically, to each of a laboratories, b materials have been sent for test and each laboratory has run each material n times. Let us suppose that the b materials cover most of the useful range of the test method under study for the type of material examined. The n determinations made by the ith laboratory on the jth material constitute what will be denoted henceforth as the "i,j cell" (see Table I). Our reasons for using this scheme will become apparent as we develop the linear model.

JOHN MANDEL, Statistician with the Division of Organic and Fibrous Materials, National Bureau of Standards, since 1947, has been engaged in research in statistical methodology, with special reference to applications in physical and chemical experimentation, and the development of test methods.

THEODORE W. LASHOF, Physicist in charge, Paper Physical Laboratory, National Bureau of Standards since 1954. Chairman of the Sampling and Conditioning Subcommittee of ASTM Committee D-6 on Paper and Paper Products and Vice-Chairman of the Precision Committee of the Technical Association of the Pulp and Paper Industry.

NOTE—DISCUSSION OF THIS PAPER IS INVITED, either for publication or for the attention of the authors. Address all communications to ASTM Headquarters, 1916 Race St., Philadelphia 3, Pa.

[1] The boldface numbers in parentheses refer to the list of references appended to this paper.

TABLE I.—INTERLABORATORY STUDY INVOLVING a LABORATORIES, b MATERIALS, AND n REPLICATIONS.

	Materials				
	No.1	No.2	No.3	...	b
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
	---	---	---	---	---
Average	---	---	---	---	---

In order to present the new basic concepts, we assume that the materials have been arranged in Table I in increasing order of the magnitude of the measurements for each material averaged over all laboratories. Now consider a graph in which the average result obtained by each laboratory for any given material is plotted against the average result of all laboratories for that material. Figure 1 shows such a graph for one laboratory. In this case, the laboratory in question agrees exactly with the average of all laboratories. Such an ideal occurrence is highly unlikely.

It is often assumed that the differences in results obtained by different laboratories are systematic in the sense that a constant systematic difference is observed between two different laboratories. If this were the case, the plot of the various laboratories against the average of all laboratories would consist of a family of parallel straight lines (Fig. 2). The fact is, however, that there exist many test methods for which the lines in question are not parallel but show changes in slope as well as vertical shifts with respect to each other (Fig. 3). Figures 4 (*a*) and (*b*), like Fig. 2, show interesting special cases of the general situation shown in Fig. 3. The linear model is based on the assumption that whereas the response lines of the various laboratories are not necessarily identical or even parallel, they nevertheless are straight lines, differing in slope or in intercept or both.

Thus, the linear model constitutes a

[1] In reference (1) the interfering factors themselves are called "I-factors." The λ-variability is then due both to scale-type errors and to the differential response of the laboratories to the I-factors.

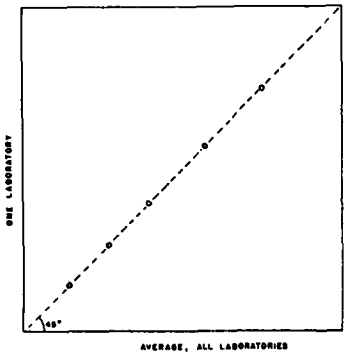


Fig. 1—Data from an ideal laboratory.

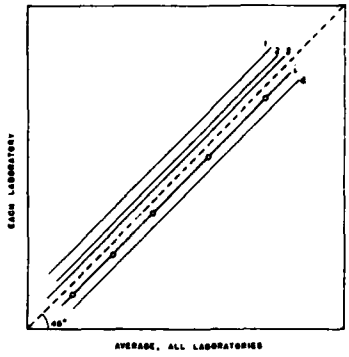


Fig. 2.—Constant systematic differences between laboratories. "Observed" material averages are shown only for the ith laboratory.

generalisation of the usual model of constant differences between laboratories. It is of course conceivable that the response lines of some of the laboratories will show curvature, requiring a second degree equation or higher to represent them. In practice, however, this situation will arise only if a laboratory is discrepant by an order of magnitude, indicating drastic departures from the prescribed procedure. When the data corresponding to such a laboratory are omitted, the remaining data conform to the linear model. But even if such a laboratory is inadvertently retained, the method of analysis proposed in this paper provides for the detection and elimination of such discrepant data.

Up to this point we have considered only the systematic differences between laboratories. Actually the observed material averages for a laboratory do not fall exactly on the line for that laboratory. This is because of within-laboratory variability. We will distinguish two types of within-laboratory variability. The first type relates to the fluctuations in results obtained on identical specimens, or if this be impossible, on specimens for which the property under study has, as closely as can be achieved, the same value. If this type of fluctuation, which we will call "replication error," ω, was the only type of within-laboratory variability, the observed averages for each laboratory, Fig. 5, could be made to fit the straight line as precisely as desired solely by increasing the number of replications.

The second type of within-laboratory variability, the effect of which cannot be reduced by merely increasing the number of replications, is less obvious. In order to illustrate its nature, let us consider the process of weighing on an analytical balance. Suppose that the weights of two samples, A and B, are to be determined, and suppose that sample A weighs a little over 1 g while sample B weighs slightly more than 5 g, thus requiring the use of two different standard weights in weighing the two samples. It is clear that the relationship between the true weights of samples A and B is known only to the extent in which two standard weights are correctly calibrated with respect to each other. The precision of this relationship cannot be improved by repeated weighings of A and B separately. Thus, consideration of more than a single sample (material) leads to a second type of within-laboratory variability, dependent on the correct relationship of the various scales, weights, or other items involved in measuring quantities of different magnitudes. This "scale-type" error can also arise from the presence of interfering substances, as in chemical analysis, or interfering properties, as in a physical method. For indeed, apart from a possible effect on the replication error, the presence of an interfering property may tend to either raise or lower the measured value, just as an improperly calibrated weight does. If different laboratories respond differently to such interfering factors, their apparent effect, in an interlaboratory study of the type here considered, will be an additional scatter of the experimental points about the straight lines corresponding to the various laboratories. This additional scatter or scale-type error, which we will refer to here as λ-variability,[1] cannot be reduced by merely increasing the number of replications.

The linear model, which we have developed here, is illustrated in Fig. 5. This figure shows a much exaggerated view of the linear systematic differences between laboratories and the within-laboratory variability of one of the laboratories. A more complete discussion of the assumptions underlying this model is given in reference (1).

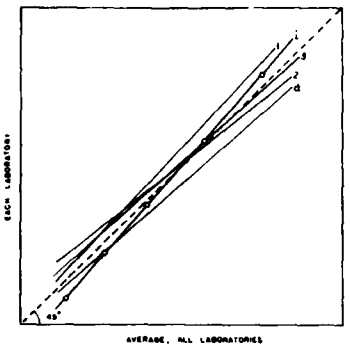


Fig. 3.—Linear systematic differences between laboratories.

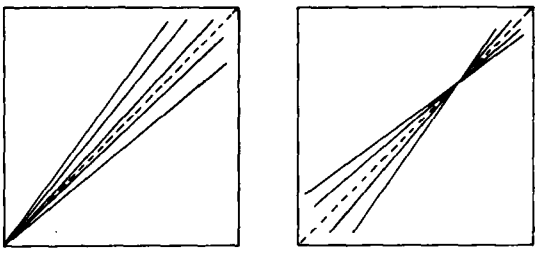
The Design of an Interlaboratory Round Robin

The linear model developed in the preceding section is based on the interlaboratory study schematically shown in Table I. This is the design which we propose. We must now fill in the details of the design.

First it is necessary to describe precisely the test method to be studied. It is surprising how vague even some widely used test methods are as regards essential details of procedure. These details should be completed through committee discussion, a survey of the literature, and experimental work within one laboratory. The draft of the detailed procedure should be circulated among all participating laboratories possibly with trial specimens, for comment and clarification.

The next question is how many and what materials are to be included in the round robin? This depends on how wide a range of materials, both as to type of material and magnitude of the property being tested, is to be covered by the test method. Also it depends on whether the instrument is a single-scale instrument or a multiple-scale instrument. Experience shows that it is desirable to use no less than five materials per scale, the five materials covering the useful range of the scale. If the study includes materials of widely different types, more materials will be needed, because in such cases, the random error will be substantially increased through the effect of λ-variability.

How many laboratories should be included? Here, a limiting factor is the amount of work involved in preparing the samples for distribution to the participating laboratories and the increase in sampling variability due to the larger amount of material required. Subject to these limitations, the number of laboratories should be as large as practicable, say, not less than 10 and preferably 20 or 30. Assurance should be obtained that each participating laboratory is properly equipped to follow all the details of procedure, and willing to assign the work to competent personnel.



(a) (b)

Fig. 4—Special cases of the general case shown in Fig. 3.

The final question to be answered before the preparation of specimens is begun is how many tests are to be run by each laboratory for each material. It is suggested that if the standard (or usual) test procedure calls for r replications, the round robin should call for an integral multiple of this number. Thus, $n = mr$ where m is an integer, preferably not less than 4. The number of replications should be as large as practicable consistent with economic considerations of time and material and statistical considerations as to the homogeneity of each material.

In making the assignments of the specimens of each material to the participating laboratories, they should be completely randomized. Of course, wherever feasible, the total portion of material used should be either selected for maximum homogeneity or, if possible, subjected to a thorough mixing prior to the assignment of specimens to the various laboratories. Any attempt to assign the specimens in such a way as to minimize within-laboratory variability at the expense of between-laboratory variability or *vice-versa* will only complicate the analysis. Experience has shown that the most satisfactory method of assignment of specimens is indeed the completely random one.

All specimens should be properly coded in such a way that only the person or persons conducting the round robin can identify the specimens. Ideally, the specimens should be thoroughly mixed so that they will be tested in random order. While this may be feasible in some cases, it may result, in many cases, in an excessive manipulation of the equipment. It is suggested, that in such cases the n specimens assigned to a given laboratory, for each material, be divided into groups such that the specimens within each group will be tested consecutively, the groups themselves being tested in a random order. For example, if 12 specimens of each of 10 materials are tested by each laboratory, one might divide the 12 specimens into 4 groups of 3 each. Each laboratory would then run the 40 groups in random order, each group consisting of 3 replicates run consecutively.

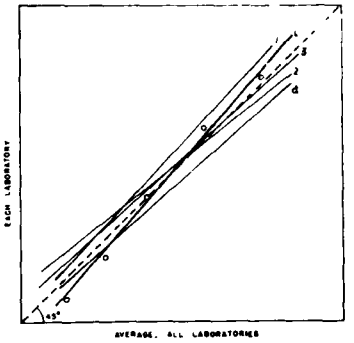


Fig. 5.—The linear model of an interlaboratory study. The observed values, including within laboratory error, are shown for the ith laboratory.

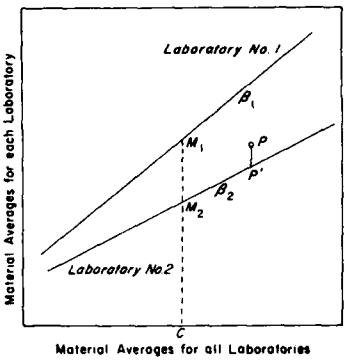


Fig. 6.—Linear model, showing the four components of variability for two laboratories. C = average of all materials and all laboratories, $M_1C = \mu_1$ = location parameter of laboratory 1, $M_2C = \mu_2$ = location parameter of laboratory 2, β_1 = slope of line for laboratory 1, β_2 = slope of line for laboratory 2, $PP' = \eta$ = departure of experimental point obtained by laboratory 2 from its response line. η comprises a component due to replication error and a component due to λ-variability.

Analysis of the Data

The purpose of the analysis is the segregation of the total error into components in accordance with the sources discussed above. Thus, we will obtain: (1) A component due to replication error; (2) a component due to λ-variability; and (3) a component due to between-laboratory variability.

In accordance with the previous discussion, the third source is expressed in terms of differences between the "response-lines" for the various laboratories. Since a straight line is determined by two parameters, the component due to between-laboratory variability will comprise two terms, corresponding to the variability of the response lines both in location and in slope. Figure 6 shows the four components of variability for two laboratories in graphical form. Each line is characterized by its location parameter, μ_i, chosen as the ordinate of the centroid and by its slope, β_i. The departure of an experimental point from the corresponding response line is composed of two component parts—the replication error and λ variability.

Analysis of the data is considered in six steps:

Step 1.—Before proceeding to the evaluation of the components of variability, it is necessary to examine the relation between replication error and the magnitude of the measurement. Table II, which relates to the Bekk smoothness data used as illustration in this paper, shows the necessity of this preliminary step. There are 14 laboratories and 14 materials, making a total of 196 cells. For each cell, the average (top figure) and the standard deviation (bottom figure) of 8 replicate measurements are given. It is quite evident that the standard deviation increases with the average. Whenever this occurs, the data are transformed into a different scale (generally of a logarithmic type) before proceeding to the subsequent steps in the analysis of the data. In this paper, we will denote values expressed in the orginal scale by the symbol y and values expressed in transformed scale by z. As a result of the transformation, the replication error ω is also transformed, and the standard deviation of the transformed replication error, which we will denote by ϵ, becomes uniform for all cells.

The formulas for the scale transformation as well as for all subsequent steps of the analysis are contained in the Appendix, in order to preserve continuity of presentation in the body of the paper.

From this point on, it will be assumed that if step 1 has indicated the need for a transformation of scale, such a transformation has been carried out on all cell averages, and that all subsequent calculations, up to and including step 5, are performed on these transformed cell averages.

If no transformation is required, all subsequent calculations are carried out on the original cell averages.

Table III(*a*) is a schematic representation of the cell averages expressed in the transformed scale, and of their row and column averages μ_i and x_j. The over-all average of all z_{ij} is denoted $\bar{z}$. The table also shows how the various parameters discussed so far and in the following steps are related by means of the equations underlying the linear

TABLE II.—BEKK SMOOTHNESS, SHOWING AVERAGES AND STANDARD DEVIATIONS IN EACH CELL.
(The column averages of these quantities are also shown. Top figure of each pair is average, bottom is standard deviation.[a])

Laboratories	Materials													
	No. 2	No. 10	No. 3	No. 4	No. 9	No. 12	No. 13	No. 5	No. 1	No. 7	No. 8	No. 6	No. 11	No. 14
No. 1	6.375	6.750	12.14	14.43	14.44	18.58	41.98	45.71	86.75	110.4	154.2	143.7	164.8	191.4
	1.14	0.556	1.55	1.59	2.37	2.46	9.25	5.70	5.99	20.1	8.83	24.7	17.3	17.8
No. 2	5.600	6.375	13.06	14.90	15.20	18.14	41.51	44.56	88.68	102.7	154.2	160.1	170.6	198.2
	0.807	1.26	2.19	1.34	1.15	1.94	4.28	5.33	13.7	25.2	14.8	34.8	14.8	16.1
No. 3	5.250	5.350	11.95	13.70	13.43	15.10	37.90	43.55	78.65	114.9	137.3	151.2	178.1	173.5
	0.754	1.39	6.23	1.33	1.53	4.31	7.46	4.74	11.2	23.2	7.49	25.3	13.2	23.7
No. 4	4.463	5.550	10.33	11.41	12.21	15.73	32.85	33.14	64.81	76.30	106.9	122.5	124.6	124.2
	0.933	0.256	1.50	0.786	0.673	1.25	5.68	2.73	10.3	5.13	8.90	19.0	10.6	20.1
No. 5	4.013	5.875	11.73	13.41	12.70	16.16	40.63	41.68	91.28	106.4	167.0	207.1	207.0	201.0
	0.681	0.167	1.94	1.04	1.61	1.54	8.51	2.30	18.5	17.0	14.7	34.7	25.1	28.0
No. 6	4.025	4.728	9.225	9.875	10.43	13.09	26.45	29.51	57.74	54.68	82.75	96.13	101.8	102.4
	0.517	0.092	1.60	0.883	0.957	0.645	5.19	1.27	2.59	8.53	7.32	16.8	15.0	16.8
No. 7	4.363	4.674	10.53	11.55	13.79	14.59	32.93	41.19	78.44	99.41	129.9	179.3	173.7	173.6
	0.709	0.301	1.75	1.31	1.48	1.70	4.61	4.26	10.2	17.0	8.75	29.7	11.4	24.3
No. 8	4.125	5.250	9.625	11.63	14.25	15.38	36.50	37.50	81.88	99.75	150.9	161.0	166.4	182.5
	0.641	0.463	1.85	1.41	1.91	1.19	4.41	4.44	10.8	15.7	16.2	24.6	14.3	27.9
No. 9	4.500	5.875	11.25	12.63	13.00	15.38	35.63	40.38	80.63	112.4	155.3	165.6	186.3	205.6
	0.535	0.354	1.04	0.744	1.51	1.41	7.13	4.14	8.63	17.9	12.6	15.4	13.8	19.6
No. 10	3.750	4.375	9.750	11.25	11.25	13.75	31.00	31.88	65.13	90.13	126.0	139.8	154.8	162.3
	0.707	0.518	1.39	0.707	1.58	1.28	7.75	4.67	5.77	19.7	14.9	23.4	12.9	17.0
No. 11	4.450	6.163	13.01	13.75	15.09	17.01	34.98	44.08	90.11	105.3	148.1	187.0	198.7	210.9
	0.737	0.457	1.88	0.750	1.93	1.81	8.04	4.52	9.88	18.8	13.5	10.4	20.2	37.9
No. 12	4.425	5.588	12.75	13.35	14.66	17.08	43.00	47.49	91.99	115.1	172.4	201.5	213.6	217.7
	0.623	0.954	1.66	1.50	2.67	2.10	5.22	7.41	14.6	36.9	22.8	31.5	11.0	28.4
No. 13	3.975	4.738	9.925	11.70	11.25	14.74	35.58	37.23	78.96	91.88	131.8	150.1	171.2	186.2
	0.434	0.250	1.51	1.80	1.43	0.955	4.61	3.17	9.74	1.68	16.6	25.0	20.2	20.3
No. 14	3.550	4.288	9.250	11.56	12.50	15.10	37.91	37.55	75.80	95.85	129.2	149.4	172.8	174.9
	0.460	0.203	1.26	1.32	1.11	1.31	8.38	3.51	9.18	18.1	9.61	15.2	14.8	15.2
Average	4.490	5.399	11.04	12.51	13.16	15.70	36.35	39.68	79.35	97.81	139.0	158.2	170.3	178.9
	0.691	0.516	1.954	1.179	1.565	1.707	6.466	4.156	10.08	17.50	12.64	23.61	15.33	22.36

[a] The Bekk smoothness data in this paper are taken from an interlaboratory study of air-leak smoothness testers conducted by a TAPPI joint Graphic Arts and Paper Testing task group.

TABLE III(a).—NOTATION FOR TRANSFORMED DATA.[a]

[a] x_{ij} = cell average, μ_i = row average, x_j = column average, and $\bar{x}$ = grand average. The basic equation for the linear model is $x_{ij} = \mu_i + \beta_i(x_j - \bar{x}) + \eta_{ij}$, where the slope β_i is further broken down according to $\beta_i = \alpha(\mu_i - \bar{x}) + \delta_i$; and the error term η_{ij} according to $\eta_{ij} = \lambda_{ij} + \sum_{k=1}^{n} \epsilon_{ijk}/n$.

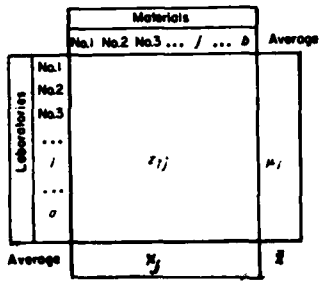


TABLE IV.—BEKK SMOOTHNESS, SHOWING ESTIMATES OF THE PARAMETERS OF THE STRAIGHT LINES CORRESPONDING TO THE VARIOUS LABORATORIES.

Laboratory	β[a]	μ	$V(\eta)$
No. 1	0.942	1601	1049
No. 2	0.962	1602	207
No. 3	0.986	1568	952
No. 4	0.912	1485	308
No. 5	1.057	1588	941
No. 6	0.883	1406	605
No. 7	1.028	1541	552
No. 8	1.028	1546	587
No. 9	1.028	1570	534
No. 10	1.020	1489	364
No. 11	1.016	1595	826
No. 12	1.055	1612	205
No. 13	1.036	1525	235
No. 14	1.048	1519	685
Average	1.000	1546	596

[a] β = slope, μ = ordinate of centroid, $V(\eta)$ = variance (fit) of points to straight line.

TABLE V(a).—ANALYSIS OF VARIANCE.

Sources of Variation	Degrees of Freedom	Sums of Squares	Mean Squares
Laboratories	$a - 1$	S_L	M_L
Materials	$b - 1$	S_M	M_M
Interaction (Laboratory × Material)	$(a - 1) \times (b - 1)$	S_{LM}	M_{LM}

TABLE V(b).—BEKK SMOOTHNESS—ANALYSIS OF VARIANCE.

Sources of Variation	Degrees of Freedom	Sums of Squares	Mean Squares
Laboratories	13	607 616	46 740
Materials	13	59 861 632	4604 741
Interaction	169	260 919	1 544

TABLE III(b).—BEKK SMOOTHNESS, SHOWING CELL, ROW, AND COLUMN AVERAGES AFTER THE DATA HAVE BEEN TRANSFORMED TO EQUALIZE THE WITHIN-CELL VARIANCES.
(The standard error of any value in the table is 19.7.)

Laboratories	Materials														Average
	No. 2	No. 10	No. 3	No. 4	No. 9	No. 12	No. 13	No. 5	No. 1	No. 7	No. 8	No. 6	No. 11	No. 14	
No. 1	804	829	1084	1159	1160	1269	1623	1660	1938	2043	2188	2157	2217	2282	1601
No. 2	748	804	1116	1173	1182	1259	1618	1649	1948	2012	2188	2204	2232	2297	1602
No. 3	720	728	1077	1137	1128	1179	1579	1639	1896	2060	2138	2180	2251	2239	1568
No. 4	650	744	1014	1057	1087	1197	1517	1520	1812	1882	2029	2088	2096	2094	1485
No. 5	603	769	1069	1127	1104	1208	1609	1620	1960	2002	2223	2316	2316	2301	1588
No. 6	605	675	965	995	1018	1117	1422	1470	1762	1739	1918	1983	2008	2010	1406
No. 7	640	670	1022	1063	1140	1164	1518	1615	1894	1997	2114	2256	2240	2240	1541
No. 8	615	720	983	1066	1154	1187	1562	1574	1913	1999	2179	2207	2221	2261	1546
No. 9	653	769	1051	1101	1114	1187	1552	1606	1906	2051	2191	2219	2270	2313	1570
No. 10	574	641	989	1051	1051	1138	1491	1503	1814	1955	2100	2146	2190	2210	1489
No. 11	648	790	1114	1138	1179	1231	1544	1644	1955	2022	2171	2272	2298	2324	1595
No. 12	646	747	1106	1126	1166	1232	1634	1677	1964	2061	2237	2304	2330	2338	1612
No. 13	599	676	997	1068	1051	1168	1551	1571	1897	1963	2120	2179	2234	2270	1525
No. 14	550	632	966	1063	1097	1179	1579	1575	1880	1982	2113	2174	2238	2243	1519
Average	647	728	1039	1095	1116	1194	1557	1594	1896	1983	2136	2192	2224	2244	1546

model. Table III(*b*) is the corresponding table for the Bekk smoothness data.

Step 2.—The second step in the analysis of the data consists in locating the straight line corresponding to each laboratory in the linear model. Mathematically, a straight line is defined by two parameters. Statistically, however, a third quantity is of interest, namely, the variance characterizing the discrepancies of the experimental points from the line representing them. Thus, for each line three quantities are of interest: the ordinate of the center of gravity of the line, μ; the slope β; and the variance $V(\eta)$, where η is the departure of an experimental point from the corresponding line. These values are computed by the usual least squares formulas for linear regression: the x values for the ith laboratory constitute the dependent variable and the averages of all laboratories for the various samples (in the transformed scale) constitute the independent variable.[3] All formulas are given in the Appendix. Table IV shows the estimated values of μ, β, and $V(\eta)$ for each laboratory calculated from the data of Table III(*b*). Note that the average of the calculated β values is, as it should be, equal to unity. The average of the μ values should be $\bar{x}$, the grand average of all values. The average of the calculated $V(\eta)$ values is an unbiased estimate of $V(\eta)$ which is needed in the next two steps.

Step 3.—At this point of the analysis, the values of $V(\eta)$ for the various laboratories should be carefully examined. If any one of these values is excessively large in comparison with the others, it is advisable to calculate the individual estimates of η, that is, the "residuals" from the regression line for the laboratory in question, in order to detect the possible presence of a completely discrepant individual point. A plot may sometimes be useful in detecting the cause of an abnormally large estimate of $V(\eta)$. In some cases, the laboratory in question may have to be omitted from the computations and a search instituted for the physical reasons of its abnormal behavior. If it is decided to omit the data for such a laboratory, the values of β, and $V(\eta)$ must be recalculated for all other laboratories.

The values of μ for the remaining laboratories are, of course, unaffected by the omission, but the over-all average $\bar{x}$ must be recomputed (see Table III(*a*)).

When the estimates of $V(\eta)$ for the individual laboratories are considered to be in satisfactory agreement, they are averaged to give an over-all estimate of $V(\eta)$. This parameter estimates the scatter of the experimental points corresponding to any given laboratory about the line for that laboratory. Part of the variability expressed by $V(\eta)$ is due to the replication error ϵ, while the remainder is precisely the λ-variability discussed in an earlier section. The partition of $V(\eta)$ into these two parts is shown in the Appendix.

Step 4.—The fourth step in the analysis of the data consists in a segregation of between-laboratory variability into two parts: the variability of the location parameter, μ, and the variability of the slope β. First, an ordinary analysis of variance is made, as indicated in Table V(*a*) and illustrated for

[3] For most purposes this procedure, which is really an approximation, will be entirely satisfactory. The reader who is interested in a more rigorous analysis will find the pertinent formulas in reference (1).

the Bekk smoothness data in Table V(b). The two variances $V(\mu)$ and $V(\beta)$ are derived from the mean-squares by formulas given in the Appendix.

The slopes and centroids of the laboratory lines may be correlated. Complete correlation occurs when all of the lines pass through a single point as in Fig. 4. In general, the correlation is not complete and $V(\beta)$ is composed of two parts, the first accounting for the correlation between μ and β, and the second for that portion of the variability of β that is unrelated to μ. This second part is denoted $V(\delta)$. The appendix gives the appropriate formulas for the partition of $V(\beta)$ into these two parts.

Step 5.—In this step the various sources of variability are considered simultaneously and the relative contribution of each source to the total variance is evaluated. In the case of nonparallel response lines for the various laboratories, the laboratory-to-laboratory component will differ with the value of the measurement (see Figs. 3 and 4). Therefore, the breakdown of variability must be evaluated separately for each region in the range over which the method is studied. In practice, it will suffice to select a few values, perhaps six in number, such that they are approximately evenly spaced over the entire range of z-values.

The components of interest are: the replication error, ϵ; the λ-variability; and the between-laboratory variability characterized by μ and β. Actually, since β is partly related to μ, the between-laboratory variability is expressible in terms of μ and δ, the latter being that part of β that is independent of μ. Therefore, a table is prepared showing, for the few selected values of z, the relative contributions of ϵ, λ, μ, and δ to the total variance of z. The first six columns of Table VI illustrate this step in the analysis of the Bekk smoothness data. Formulas for this step are given in the Appendix.

Step 6.—Finally, in case a transformation of scale was required, the total variance $V(z)$, or rather its square root, the standard deviation of z, is converted back into the original scale, giving σ_y. It is also useful, in this case, to convert the values of z chosen for the calculations in Table VI, into corresponding y values, that is, into the original scale. The last two columns in Table VI illustrate this step for the Bekk smoothness data.

TABLE VI.—BEKK SMOOTHNESS, SHOWING RELATIVE IMPORTANCE OF THE VARIOUS SOURCES OF VARIABILITY.

z-Scale		Sources of Variability				y-Scale	
Average z	Total Variance $V(z)$	Within Laboratory		Between Laboratory		Average, sec y	Standard Deviation σ_y
		$\frac{V(\epsilon)}{V(z)}$	$\frac{V(\lambda)}{V(z)}$	$\frac{[1 + \alpha(z - \bar{z})]^2 V(\mu)}{V(z)}$	$\frac{(z - \bar{z})^2 V(\delta)}{V(z)}$		
2224	10 011	0.31	0.02	0.57	0.10	170 3	39.2
1896	8 028	0 38	0.03	0.56	0.03	79.4	16.4
1557	6 636	0.47	0.03	0.50	0.00	36.4	6 8
1116	5 814	0.53	0.04	0.36	0.07	13.16	2 3
647	6 169	0.50	0.03	0.18	0.28	4 49	0.81

Interpretation of the Analysis

In interpreting the results of the analysis, the points of major interest are (*a*) the relative importance of the various sources of error, (*b*) the steps required to improve precision, if necessary, and (*c*) the need for standard samples. For these purposes, the following procedure is recommended.

1. Compare $V(\lambda)$ and $V(\epsilon)$: the relation between these two quantities will reveal how much can be expected from mere replication of measurements. If $V(\lambda)$ is large with respect to $V(\epsilon)$, replication is generally a waste of time. Even if $V(\lambda)$ is smaller than $V(\epsilon)$, replication is useful only to the point of making $V(\epsilon)/n$ small with respect to $V(\lambda)$. Thus, in the case of the Bekk smoothness data (Table VI) the replication error exceeding $V(\lambda)$ by a factor of ten, approximately, an effective increase in precision will result from ten replications. But a number of replications considerably larger than ten would be wasteful since the limiting factor, at that point, is $V(\lambda)$ which is unaffected by replication.

2. Study the table of values of β and μ for the various laboratories. Occasionally, a single laboratory (or a small group of laboratories) is discrepant in one or both these parameters, while all others are in close agreement. An investigation of the causes of such discrepancies is then indicated, and the analysis of variance carried out by omitting the discrepant laboratory (or laboratories) may be more meaningful than that based on its inclusion.

In Table IV, two of the laboratories show much smaller values for both β and μ than the other laboratories. These two were foreign laboratories where the standard relative humidity is appreciably higher than in the American laboratories. However, even this appreciable difference in procedure does not require the omission of these laboratories when the analysis is made using the linear model.

3. Compare the total between-laboratory variability for various values of z with the within-laboratory variability (Table VI) keeping in mind that the effect of $V(\epsilon)$ will depend on the number of replications which is called for by the standard method. If the total between-laboratory variability is small compared with the within-laboratory variability throughout the table, all of the laboratories are essentially in agreement and only refinement of the procedure to reduce $V(\lambda)$ can improve the precision of the method.

If $V(\lambda)$ is so large that the method is not sufficiently precise to be useful, the possible cause of a large $V(\lambda)$ should be investigated. Perhaps types of materials were included in the round robin for which the method was not designed. Perhaps the method as written fails to call for the control of important interfering conditions or fails to correct for significant interfering properties.

If the between-laboratory variability is not negligible, examine its two terms separately. If the term in $V(\delta)$ may be neglected, the lines will form a simple pattern: they will either converge to a point (or a small region) or be, for all practical purposes, parallel. In either case, the calibration of the method at a single point other than the point of convergence will suffice to obtain the maximum possible agreement among the laboratories. On the other hand, if the term in $V(\delta)$ becomes appreciable anywhere in the table, the lines for the different laboratories will tend to criss-cross at random and the method will require calibration at two points. (This is the situation for the Bekk smoothness data of Table VI.) There is one exception: if the term in $V(\mu)$ is negligible but not the term in $V(\delta)$, the lines just happen to converge at the centroid, and calibration will be required at a single point as far away from the centroid as is practical.

In general, the term in $V(\mu)$ will not be negligible throughout the table. If the variation in this term is small, the place of the required calibration point or points in the range of the measured quantity is immaterial, except that when two points are required they should be located as far apart as practical. If the variation in the $V(\mu)$ term is appreciable (as for the Bekk smoothness data shown in Table VI), the lines will partially or completely converge and the calibration point or points should be located to avoid the area of convergence.

In summary, if between-laboratory variability is greater than within-lab-

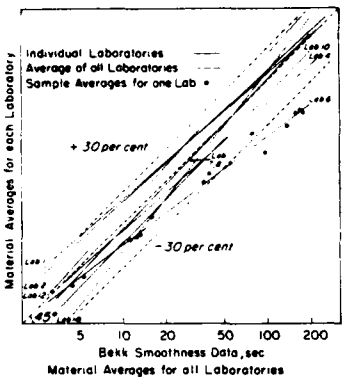


Fig. 7.—Bekk smoothness data, showing nonparallelism of the laboratory lines. The scale on both coordinate axes is logarithmic, but the vertical deviation of each point from the 45 deg line has been doubled in order to clearly show the differences between the laboratories. The broken lines correspond to deviations of plus or minus 30 per cent from the average (45 deg line). A few of the laboratories close to the 45 deg line have been omitted for clarity. The material averages are shown for one laboratory for which the fit of the points to the line is typical.

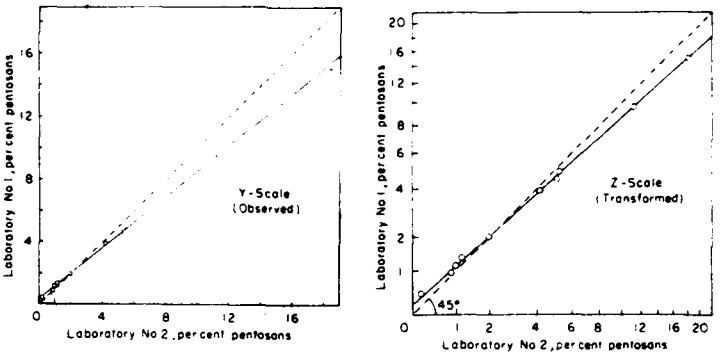


Fig. 8—Pentosans, by orcinol, showing comparison for two laboratories. The slope of the line is distinctly different from unity in either scale. Therefore, these two laboratories will show nonparallelism in a graph of the type shown in Fig. 3. (These data are from an interlaboratory study of orcinol, aniline acetate and bromination methods for the determination of pentosans in pulps. This study was under the direction of a joint ACS-ASTM-TAPPI-ICCA task group.)

oratory variability and greater than can be tolerated for practical application of the method, the method must have better standardization. This can be done by using one or two standard samples to calibrate the method at appropriately chosen values of the measured quantity.

Comparison with Other Models

In the previous sections we have developed the linear model for the measuring process and discussed the design, analysis, and interpretation of an interlaboratory study in accordance with this model. The question naturally arises as to how this model compares with the models underlying the more conventional statistical designs and analyses.

The simplest interlaboratory study is one in which a random sample of a particular material is sent to each of two or more laboratories and they are asked to report the value of some property of the material. If the laboratories turn in values that are in satisfactory agreement with each other, the methods used by the laboratories are considered to be satisfactory and everyone is happy. If the values do not agree within the hoped-for limits, this simple study is unable to furnish even the slightest hint as to the cause or causes of the unsatisfactory results.

Had each laboratory been asked to make duplicate or replicate measurements on the sample it received, the differences found between laboratories might be found to be entirely accounted for by the inability of each laboratory to duplicate its own results. While the simple study fails to distinguish between the variability within and between laboratories, the second type is generally interpreted in terms of a model that allows for random variability between laboratories beyond the within-laboratory fluctuation. This is the model which, with some ramifications, is very frequently used in interlaboratory studies of test methods.

A typical ramification is to have several analysts in each laboratory. Also each analyst may repeat the test on each of several days. The result is an hierarchical or nested design (2, p. 884f) which provides information on the relative importance of the various possible sources of within-laboratory variability. Very often each analyst is also asked to make determinations by each of two or more methods. This gives a two-way or cross design which may or may not be nested each way. Obviously three- or more-way crosses could be (and undoubtedly have been) used.

Interlaboratory studies sometimes use two or three materials. Even in the case where each laboratory makes only one determination per material, the use of more than one material per laboratory provides information on laboratory "biases." The analysis of data of this type is usually made in accordance with conventional two-way analysis of variance procedures (2, p. 888f; 3), but an ingenious graphical method has recently been developed (4). Both of these methods require that the within-laboratory test error be the same for all laboratories and materials. If this is not the case, the analysis of the data has often been carried out separately for each material, in accordance with the usual "between-within" type of analysis of variance (3, 5).[4] However, in cases in which the standard deviation of the within-laboratory error is a known function of the magnitude of the measurement, an appropriate transformation of the data prior to analysis will ensure a homogeneous error term and permit a two-way analysis of variance. In particular, the simple logarithmic transformation is often used (6, pp. 116, 137); it is based on the assumption of a constant coefficient of variation (error proportional to magnitude) for within-laboratory test error. The study which we report in this paper, Bekk smoothness, is an example of a proportional type of error. In many cases a straight-line relationship with nonzero intercept, rather than a simple proportionality, is found (see Eq 1) and, hence, the transformation given in Eq 2 is required.

If an interlaboratory test has been run in accordance with a two-way classification with replications within cells, such as shown in Table I, it is usually interpreted on the basis of a model allowing for constant laboratory differences ("biases") and, in the case of a significant interaction term, for an additional random "variable bias" (6, p. 124). According to such a model, the response lines of all laboratories are necessarily parallel to each other, except for random scatter, and a plot of the results of one laboratory *versus* those of another is a straight line of 45 deg slope. This follows from the assumption that the "variable bias" is a random effect. It appears, therefore, merely as additional scatter about the 45-deg line.

There is, however, considerable evidence for the existence of nonconstant, nonrandom differences between labora-

[4] Tentative Recommended Practice for Interlaboratory Testing of Textile Materials (D 990 - 54 T), 1958 Book of ASTM Standards, Part 10.

tories. The Bekk smoothness data of Table II are shown graphically in Fig. 7. The nonparallelism of the lines is quite evident. The Bekk smoothness test is a physical test. However, the same type of results has been obtained with chemical tests. Figure 8 shows the relationship between the results obtained by two laboratories in the determination of pentosans in a series of pulp samples using a colorimetric method. Despite the fact that each laboratory prepared a calibration curve of the intensity of the color in terms of samples of known composition, the relation between the results of the two laboratories is definitely not the expected straight line of slope one and passing through the origin. Nor does a constant bias for each laboratory explain their relationship. It is seen that while a straight line is an adequate representation of this relation, this line has a slope distinctly different from unity, in addition to a nonzero intercept.

The conventional model for two-way classification data has a further disadvantage when applied to interlaboratory data. It is extremely sensitive to "outlying" data. Even a single outlier may result in a considerably enlarged interaction term. The practice of eliminating outliers on the basis of control-chart procedures has often led to the discarding of a substantial proportion of the participating laboratories. It is probably for these reasons that interlaboratory studies have been considered to be so "frustrating" (7).

The model presented in this paper allows for nonconstant, nonrandom differences between laboratories, by allowing their response lines to have slopes different from 45 deg, in addition to nonzero intercepts. It safeguards against the effect of outliers in two ways: by a preliminary analysis of the relation between within-cell variability and cell average, and by a separate calculation, for each individual laboratory, of the scatter of its points about its response line. It is believed that this model provides an adequate basis for the general description of the precision of measuring processes and a satisfactory procedure for analyzing and interpreting interlaboratory studies.

References

(1) John Mandel, "The Measuring Process," *Technometrics*, Aug., 1959 (in press).

(2) M. D. Finkner, "The Reliability of Collaborative Testing for A.O.A.C. Methods," *Journal* Assn. Official Agricultural Chemists, Vol. 40, No. 3; Aug., 1957, pp. 882–892.

(3) W. G. Cochran and G. M. Cox, *Experimental Designs*, John Wiley & Sons, Inc., New York, N. Y., Chapter 14 (1957).

(4) W. J. Youden, "Statistical Design," *Industrial and Engineering Chemistry*, "Presentation for Action," Vol. 50, No. 8, Aug., 1958, pp. 83*A*, 84*A*; "Product Specifications and Test Procedures," Vol. 50, No. 10, Oct., 1958, pp. 91A, 92A; "Circumstances Alter Cases," Vol. 50, No. 12, Dec., 1958, pp. 77A, 78A; "What is a Measurement?," Vol. 51, No. 2, Feb., 1959, pp. 81A, 82A.

(5) D. S. McArthur, *et al.*, "Evaluation of Test Procedures," *Analytical Chemistry*, Vol. 26, No. 6, June 1954, pp. 1012–1018.

(6) Owen L. Davies, *Design and Analysis of Industrial Experiments*, Hafner Publishing Co., New York, N. Y. (1954).

(7) C. A. Hochwalt, "Standards and the New Science of Materials," "Tomorrow's Standards—Equations of State," ASTM Bulletin, No. 230, May, 1958, p. 29(TP107)

Appendix

Computations

The computations are set forth in steps which are numbered identically as in the body of the paper.

Step 1. Scale Transformation

Compute, for each cell of Table I, its average and its standard deviation. Denote the average of the i,jth cell by y_{ij} and its standard deviation by s_{ij}. Then for each material (that is, column) compute the average $\bar{y}_j$ of the cell averages y_{ij} and the average $\bar{s}_j$ of the standard deviations s_{ij}. The results are shown in Table II for the Bekk smoothness data. Prepare a graph plotting the standard deviation $\bar{s}_j$ *versus* the average $\bar{y}_j$, as in Fig. 9, and fit a simple curve to the points thus obtained. For the data of Fig. 9, a straight line through the origin is a good fit. In general, a straight line, not necessarily passing through the origin, will be sufficient, as only an approximate, order-of-magnitude relationship is required. Determine the intercept A and the slope B of this straight line, in accordance with the equation:

$$\bar{s}_j = A + B\bar{y}_j + (\text{random fluctuation}) \quad (1)$$

For the Bekk smoothness data, this equation is given by $A = 0$ and $B = 0.128$.

If the slope B is appreciably different from zero, use the transformation

$$z = K \log (A + By) - C \quad \ldots(2)$$

where K and C are arbitrary constants, the values of which are chosen on the basis of convenience. Theoretically, the transformation should be applied to each of the n observations in each cell. In most cases, however, it is sufficient to apply the transformation to the cell averages. Table III(a) shows schematically the transformed averages z_{ij} for each cell, and Table III(b) shows the transformed averages for the Bekk smoothness data. Since, for these data, we found $A = 0$, Eq 2 becomes, in this case,

$$z = K \log y - (C - K \log B)$$

It was convenient to make $K = 1000$ and $C - K \log B = 0$. Thus, the transformation, here, is simply $z = 1000 \log y$.

As a result of the transformation, the error ω has been transformed into a different error, denoted by ϵ, the variance of which is constant for all cells in Table III. Its value is given by the following expression:[4]

$$V(\epsilon) = \left(\frac{KB}{2.3}\right)^2 \quad \ldots(3)$$

For the Bekk smoothness data we find

$$V(\epsilon) = \left(\frac{1000 \times 0.128}{2.3}\right)^2 = 3097$$

Since each cell contained 8 replicates, the standard error of a cell average is $\sqrt{3097/8} = 19.7$.

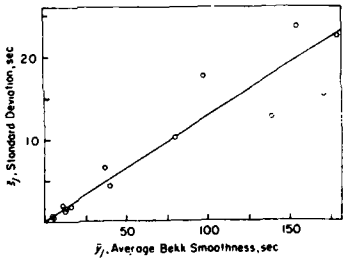


Fig. 9.—Bekk smoothness data, showing approximate linear relationship between standard deviation and magnitude of the smoothness value. The straight line is: $\bar{s}_j = 0.128\ \bar{y}_j$. Thus $A = 0$ and $B = 0.128$.

If the slope B is not appreciably different from zero, no transformation of scale is required and all subsequent steps are carried out on the original cell averages. In this case, $y = z$ and $\omega = \epsilon$.

Step 2. Determination of μ_i, β_i, and $V_i(\eta)$

First compute the row and column averages in Table III(a) as follows:

$$\mu_i = \frac{1}{b} \sum_j z_{ij} \quad \ldots(4)$$

$$x_j = \frac{1}{a} \sum_i z_{ij} \text{ and } \bar{x} = \frac{1}{b} \sum_j x_j \quad \ldots(5)$$

[4] The numerical factor 2.3 is due to the use of logarithms to the base 10 and equals $\log_e 10$.

Then compute the quantities below in the indicated order:

$$X = \sum_j x_j^2 - b\bar{x}^2 \quad \ldots\ldots(6)$$

$$Z_i = \sum_j z_{ij}^2 - b\mu_i^2 \quad \ldots\ldots(7)$$

$$P_i = \sum_j x_j z_{ij} - b\mu_i \bar{x} \quad \ldots\ldots(8)$$

$$\beta_i = \frac{P_i}{X} \quad \ldots\ldots(9)$$

$$V_i(\eta) = \frac{a}{a-1} \cdot \frac{1}{b-2}\left[Z_i - \frac{P_i^2}{X}\right] \quad \ldots(10)$$

The last formula constitutes a slight departure from the ordinary calculations in linear regression: the correction factor $a/(a-1)$ is due to the fact that the errors of z_{ij} and x_j are slightly correlated.

The results of these computations for the Bekk smoothness data are shown in Table IV.

Determination of $V(\mu)$ and $V(\lambda)$

To obtain $V(\eta)$, average all values of $V_i(\eta)$ (see Table IV). $V(\lambda)$ is given by the following equation:

$$V(\lambda) = V(\eta) - (1/n)\ V(\epsilon) \quad \ldots(11)$$

where $V(\epsilon)$ is given by Eq 3. Should the estimate of $V(\eta)$ be less than that of $(1/n)\ V(\epsilon)$, then $V(\lambda)$ must be considered to equal zero. For the Bekk smoothness data, we obtain $V(\lambda) = 596 - (3097/8) = 209$.

Step 3. Determination of Variances

Table V(a) is constructed in the usual manner, using the transformed cell averages. Note that $S_M = aX$ where X is given by Eq 6, and that

$$S_{LM} = \left(\sum_i Z_i\right) - S_M = \left(\sum_i Z_i\right) - aX \quad \ldots(12)$$

The variances $V(\mu)$ and $V(\beta)$ may now be obtained from the previously obtained value of $V(\eta)$ and from Table V(a).

$$V(\mu) = \frac{M_L - V(\eta)}{b} \quad \ldots\ldots(13)$$

$$V(\beta) = \frac{A[M_{LM} - V(\eta)]}{M_M - V(\eta)} \quad \ldots(14)$$

If either of these equations yields a negative value, the corresponding variance is taken to be zero.

The quantity $V(\delta)$ is obtained from the following equation:

$$V(\delta) = V(\beta) - \alpha^2 V(\mu) \quad \ldots\ldots(15)$$

where $V(\beta)$ is given by Eq 14 and α by the equation:

$$\alpha = \frac{ab\left[\sum_i \mu_i P_i - \bar{x}\sum_i P_i\right]}{S_L S_M} \quad \ldots(16)$$

For the Bekk smoothness data, these computations yield:

$V(\mu) = 3296$ $\quad V(\beta) = 0.002881$ $\quad \alpha = 0.0004661$

$V(\delta) = 0.002881 - 0.000716 = 0.002165$

Step 4. Breakdown of Total Variance

The total variance of z, for any value of z, is given by the equation:

$$V(z) = V(\epsilon) + V(\lambda) + [1 + \alpha(z - \bar{x})]^2 V(\mu) + (z - \bar{x})^2 V(\delta) \quad \ldots(17)$$

From this equation we derive, by dividing by $V(z)$, the breakdown of the total variance into its fractional parts:

$$1 = \frac{V(\epsilon)}{V(z)} + \frac{V(\lambda)}{V(z)} + \frac{[1 + \alpha(z - \bar{x})]^2 V(\mu)}{V(z)} + \frac{(z - \bar{x})^2 V(\delta)}{V(z)} \quad (18)$$

This information is tabulated in Table VI for the Bekk data, for values of z corresponding to some selected values of y covering the range of interest.

Step 5. Conversion to Original Scale

The total variance of z is converted back to the y-scale by means of the equation:

$$V(y) = \left[\frac{2.30}{K}\left(\frac{A}{B} + y\right)\right]^2 V(z) \quad (19)$$

It is generally desirable to add two more columns to the table showing the breakdown of $V(z)$: a column of the selected values for which the breakdown was made and a column of σ_y, obtained by taking the square root of Eq 19 (see Table VI).

ASTM BULLETIN

Authorized Reprint from the Copyrighted ASTM BULLETIN No. 239, July, 1959
Published by the American Society for Testing Materials, Philadelphia 3, Pa.

Journal of Research of the National Bureau of Standards Vol. 53, No. 3, September 1954 Research Paper 2527

Sensitivity—A Criterion for the Comparison of Methods of Test

J. Mandel and R. D. Stiehler

In the evaluation of many methods of test, the two usual criteria—precision and accuracy—are insufficient. Accuracy is only applicable where comparisons with a standard can be made. Precision, when interpreted as degree of reproducibility, is not necessarily a measure of merit, because a method may be highly reproducible merely because it is too crude to detect small variations.

To obtain a quantitative measure of merit of test methods, a new concept—sensitivity—is introduced. If M is a measure of some property Q, and σ_M its standard deviation, the sensitivity of M, denoted ψ_M, is defined by the relation $\psi_M=(dM/dQ)/\sigma_M$. It follows from this definition that the sensitivity of a test method may or may not be constant for all values of the property Q. A statistical test of significance is derived for the ratio of sensitivities of alternative methods of test. Unlike the standard deviation and the coefficient of variation, sensitivity is a measure of merit that is invariant with respect to any functional transformation of the measurement, and is therefore independent of the scale in which the measurement is expressed.

1. Introduction

In the physical sciences, there frequently is a choice between several methods for the determination of a particular characteristic. In such cases means are necessary to compare the relative merits of the various methods. The customary procedure for evaluating a test method, particularly in analytical chemistry, is to determine accuracy by comparing the values found on known samples with the theoretical values, and to express precision by the reproducibility of the experimental values as measured by the standard deviation. Alternative methods can then be compared on the basis of both precision and accuracy. In the evaluation of many methods of test, particularly those for polymeric materials, these criteria are insufficient. This paper presents a single criterion by which the relative merit of methods of test can be evaluated. The main advantage of the new criterion—referred to as sensitivity—is that it takes into account, not only the reproducibility of the testing procedure, but also its ability to detect small variations in the characteristic to be measured.

The need for such a criterion has been felt by various workers. Newton [1] [1] discusses the fallacy of comparing alternative test methods on the sole basis of their respective standard deviations of error. According to Throdahl [2], Mooney considers a *coefficient of discrimination*, defined as the ratio of the difference between the average values obtained from two sets of samples to the standard deviation within samples. Dillon [3] compares two plastometers on the basis of their *selectivities*, the concept of selectivity being defined by him as the "percentage difference between two observations on different mixtures divided by the average maximum percentage error." Roth and Stiehler [4], in comparing the precisions of strain and stress measurements, convert the standard deviation of strain into stress units and then consider the ratio of this converted standard deviation to that of stress; alternatively, they consider the ratio of the variance "between batches" to that "within batches" as a criterion for the sensitivity of either method. The latter criterion is also applied by Buist and Davies [5] and by Newton, Scott, and Whorlow [6], who refer to it as the *discriminating power*. Reichel [7] introduces the concept of *"technische Güte"* to characterize the merit of methods of chemical analysis.

In this paper, a general mathematical definition is proposed for the *sensitivity concept*, which is an intrinsic measure of merit, of particular value for the comparison of two or more alternative test methods.

2. Sensitivity in the Case of Proportionality

In most analytical methods in chemistry the desired material is not determined directly but is calculated from measurements of a proportional quantity of some related material. For example, in the determination of zinc, the amount of this metal is calculated from the quantity of zinc oxide, zinc sulfate, or other zinc compound actually measured. In comparing the relative merits of the use of these alternative compounds, a pertinent consideration, besides the magnitude of experimental error, is the ratio of the equivalent weight of the zinc compound to that of zinc. It is recognized that a larger ratio is preferable, provided that the experimental error is not increased in the same proportion. A correct evaluation of alternative methods, involving zinc compounds of different equivalent weight, can be obtained from the following considerations:

[1] Figures in brackets indicate the literature references at the end of this paper.

The percentage of zinc in the unknown is given by the equation

$$\text{Zn}=\frac{100P}{W}\times\frac{[\text{Zn}]}{[\text{Zn compound}]}, \tag{1}$$

where P is the weight of the Zn compound measured; W is the weight of the sample; [Zn] is the equivalent weight of zinc; and [Zn compound] is the equivalent weight of the zinc compound measured.

Let Q equal the percentage of zinc, R the ratio of the equivalent weights of zinc and the zinc compound measured, and M the weight of zinc compound per gram of sample. Then

$$Q=100MR. \tag{2}$$

From this relation it follows [8] that the standard deviation for the determination of zinc is given by the equation

$$\sigma_Q=100R\sigma_M. \tag{3}$$

Equation (3) shows that the precision of the zinc determination is improved when (1) the quantity $100R$ is small, and (2) the error of measurement of the zinc compound (σ_M) is small.

If the weight of zinc compound per gram of sample is plotted against the percentage of zinc, a straight line is obtained, as shown in figure 1. The line passes through the origin and has a slope equal to the reciprocal of $100R$. Let the slope be designated as K. Equation (3) can now be written

$$\sigma_Q=\frac{\sigma_M}{K}. \tag{4}$$

Thus, high precision in the determination of Q (i. e., a small value for σ_Q) reduces to the requirement that the quantity K/σ_M be large. The absolute value of the quantity K/σ_M is defined as the sensitivity of the measurement of M for the determination of Q and is denoted by ψ. Thus

$$\text{Sensitivity}=\psi=\frac{|K|}{\sigma_M}. \tag{5}$$

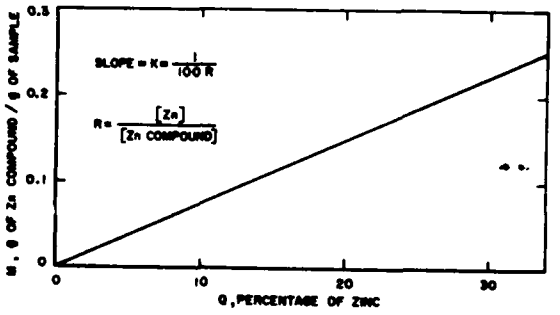


FIGURE 1. *Sensitivity for proportional relationship.*

It is obvious that the merit of the method is dependent on more than the reproducibility of measurement of M. It also depends on the rate of change in M with a change in Q or the ability to discriminate between small changes in Q.

3. Sensitivity in the General Case

In many methods, particularly when dealing with polymeric materials, the measured quantity M and the desired quantity Q are not linearly related. An example is the measurement of refractive index to determine the percentage of bound styrene in GR-S synthetic rubber. Additional difficulties arise when it becomes impossible to define a single criterion Q for the characterization of the properties in which one is interested. In these cases it is necessary to consider a measurable quantity M that is in some sense related to these properties. An example of this type is given by vulcanization tests on rubbers, where stress-strain measurements are used as an index or measure of the degree of vulcanization. Whether or not a quantity Q can be defined, and whatever the relation may be between a characteristic Q and the measured quantity M, the criterion defined as sensitivity can effectively be used for evaluating and comparing methods of test.

Figure 2 illustrates a case in which Q is susceptible of exact definition and the relation between M and Q is curvilinear. If it is desired to differentiate between the two close values, Q_1 and Q_2, by means of the corresponding measurements M_1 and M_2, it is again apparent that the success of the operation will depend on two circumstances: (1) the magnitude of the difference M_2-M_1, for a given difference Q_2-Q_1; i. e., the magnitude of the slope $(M_2-M_1)/(Q_2-Q_1)$; and (2) the precision of measurement; i. e., the smallness of the standard deviation. Indeed, if σ_M is too large, the regions of uncertainty of M_1 and M_2 may overlap, and the discrimination fail. As before, these two desiderata can be combined in a single criterion, the sensitivity, defined according

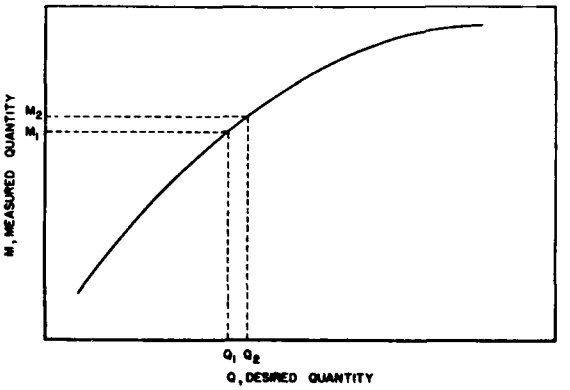


FIGURE 2. *Sensitivity for curvilinear relationship.*

to eq (5) as the absolute value of the ratio of the slope $K=(M_2-M_1)/(Q_2-Q_1)$ to the standard deviation of M, σ_M. The larger the sensitivity, the more useful will be the test method M for the characterization of Q. It should be noted, however, that in the general case, K is no longer constant but varies with the value of Q. Thus, even in cases in which the experimental error (measured by σ_M) remains constant, the sensitivity may vary with the value of Q. Only when the error is proportional to K is the sensitivity constant.

If the properties under consideration cannot be expressed by means of a single criterion Q, it is not possible to determine the absolute sensitivity of a method of test. It is possible, however, to determine the relative sensitivities of two or more methods used to characterize these properties. This important application of the sensitivity concept can best be shown by first considering a case in which a single criterion Q exists, and two alternative measuring methods M and N, both related to Q, are to be compared. For example, density and refractive-index methods for determining the bound styrene in GR-S may be compared without knowing the actual percentage of bound styrene. Let ψ_M and ψ_N be the sensitivities corresponding to the two methods. From eq (5) it follows that the ratio of the sensitivities is given by

$$\frac{\psi_M}{\psi_N}=\frac{|K_M/K_N|}{\sigma_M/\sigma_N}=\frac{|K'|}{\sigma_M/\sigma_N}. \tag{6}$$

The meaning of K' is found as follows:

$$K'=\frac{K_M}{K_N}=\frac{\Delta M/\Delta Q}{\Delta N/\Delta Q}=\frac{\Delta M}{\Delta N}. \tag{7}$$

Thus K' is the slope of a curve of M plotted as a function of N. From eq (5) it follows that the dimension of sensitivity is that of $1/Q$, since σ_M has the dimension of M, and K is of dimension M/Q. On the other hand, the ratio of the sensitivities of alternative test methods given in eq (6) is dimensionless. This fact, as well as eq (7), shows that the comparison of two methods, by means of the ratio of their sensitivities, does not necessitate a knowledge of their relation to the theoretical Q. All that is required is a knowledge of their mutual relationship.

In the case of bound styrene, the relation between density and refractive index can be established from a series of samples of different bound styrene contents without a knowledge of bound styrene in any sample. Of course, the bound styrene content could be determined by some absolute method, and the absolute sensitivities of the refractive index and density methods for measuring this property could be established.

In the case of stress–strain measurements, on the other hand, the characteristic—degree of vulcanization—cannot be represented by a single quantity Q and consequently no absolute sensitivities for either method can be calculated. Nevertheless, relation (6), with K' given by (7), can be applied, since it does not involve the quantity Q, and the sensitivity ratio can be used to compare the measurement of tensile stress [9] and the measurement of strain [4]. The relationship between these two methods of measurement for a GR–S synthetic rubber compound, according to Roth and Stiehler [4], is given by the equation:

$$SE^n=C \tag{8}$$

where S represents tensile stress, E represents strain, and n and C are constants for any particular type of vulcanizates.

If the logarithmic derivative is taken, it follows that

$$\frac{dS}{S}=-n\frac{dE}{E}. \tag{9}$$

As n is of the order of 1.5, it might be expected that measurements of tensile stress would detect variations in the vulcanizates better than measurements of strain. However, Roth and Stiehler [4] show that the error of measurement of strain is much smaller than that of the usual measurement of tensile stress; hence, the sensitivity of strain measurements is greater.

From eq (9) it follows that the slope of the strain versus tensile-stress curve is

$$\frac{dE}{dS}=-\frac{E}{nS},$$

and consequently,

$$\frac{\psi_E}{\psi_S}=\frac{E/nS}{\sigma_E/\sigma_S}. \tag{10}$$

This expression is found to exceed unity, as shown in table 1, which lists data pertinent for the calculation of the sensitivity ratio, for tensile-stress and strain values obtained in three different plants and for two cures [10]. It should be noted that the ratio of the two sensitivities varies with the degree or time of cure, since the factor E/nS decreases as vulcanization progresses. The advantages of the strain test are therefore greatest for tests on vulcanizates that are undercured. The data also show that the greater sensitivity of the strain test is due to its better reproducibility.

TABLE 1. *Comparison of tensile stress and strain measurements of GR-S synthetic rubber*

Cure at 292° F	Plant	K' $(E/1.6\,S)$ *	Standard deviation		Ratio of sensitivities (strain/stress)
			Strain at 400 psi	Stress at 300% elongation	
min		% *psi*	%	*psi*	
25	A	0.610	1.6	9.5	3.6
	B	.542	3.1	22.5	3.9
	C	.362	2.1	15.4	2.6
100	A	.0706	0.83	14.8	1.3
	B	.0703	1.84	35.8	1.4
	C	.0641	1.17	37.1	2.0

* The value 1.6 taken for n is an upper limit for GR–S synthetic rubber. For values of n smaller than 1.6, the ratios in the last column will be larger.

relating them are not linear. An important advantage of the sensitivity concept is its nondependence on the scale of measurement. The standard deviation, being expressed in the same units as the measurement, has a value that depends on the unit and scale in which the measurement is expressed. The coefficient of variation, which is defined as the ratio of the standard deviation to the mean value, is nondimensional, because both these quantities are expressed in the same units. However, except for scales that are proportional to each other, the coefficient of variation is dependent on the scale in which the measurement is expressed.

Consider, for example, the logarithmic transformation of a measurement y:

$$z=\ln y.$$

The standard deviation of z is then approximated [8] by the expression

$$\sigma_z=\frac{d \ln y}{dy}\sigma_y=\frac{\sigma_y}{y}.$$

It is evident, from this formula, that the coefficient of variation of z, σ_z/z, is in general different from that of y, σ_y/y. It can be shown that the only transformation that leaves the coefficient of variation rigorously unaltered is a proportional transformation: $z=ky$, i. e., a simple change of units. (To the extent that the approximate expression $\sigma_z=|dz/dy|\sigma_y$ is applicable—[for details see 12, secs. 27.7 and 28.4]—the coefficient of variation is also unaltered under the transformation $z=k/y$.)

On the other hand, the sensitivity of the transformed variable z, for *any* transformation

$$z=f(y) \tag{12}$$

is identical to that of the original variable y, to the extent that the following calculation of the ratio of the two sensitivities is applicable:

$$\frac{\psi_z}{\psi_y}=\frac{\left|\frac{dz}{dy}\right|}{\sigma_z/\sigma_y}=\frac{\left|\frac{dz}{dy}\right|}{\left|\frac{dz}{dy}\right|\sigma_y/\sigma_y}=1. \tag{13}$$

It is evident from eq (13) that sensitivity is not affected by any transformation of the measurement, and is therefore independent of the scale in which the measurement is expressed.

6. References

[1] R. G. Newton, Proc. Second Rubber Technol. Conf., p. 233, Institution of the Rubber Industry, London (W. Heffer & Sons, Ltd., Cambridge, England, 1948).
[2] M. C. Throdahl, J. Colloid Sci. **2,** 187 (1947); Rubber Chem. and Technol. **21,** 164 (1948).
[3] J. H. Dillon, Physics **7,** 73 (1936); Rubber Chem. and Technol. **9,** 496 (1936).
[4] F. L. Roth and R. D. Stiehler, J. Research NBS **41,** 87 (1948) RP1906; India Rubber World **118,** 367 (1948); Rubber Chem. and Technol. **22,** 201 (1949).
[5] J. M. Buist and O. L. Davies, Trans. Inst. Rubber Ind. **22,** 68 (1946); Rubber Chem. and Technol. **20,** 288 (1947).
[6] R. G. Newton, J. R. Scott, and R. W. Whorlow, Proc. Intern. Rheol. Congr. **II,** 204; **III,** 61 (1948).
[7] E. Reichel, Z. anal. Chem. **109,** 385 (1937).
[8] W. E. Deming, Statistical adjustment of data, ch. III (John Wiley & Sons, Inc., New York, N. Y., 1943).
[9] ASTM Book of Standards, pt. 6, D-412-51T, p. 134 (American Society for Testing Materials, Philadelphia, Pa., 1952).
[10] Private communication from the Office of Synthetic Rubber, Reconstruction Finance Corp.
[11] R. L. Anderson and T. A. Bancroft, Statistical theory in research (McGraw-Hill Book Co., New York, N. Y., 1952).
[12] H. Cramer, Mathematical methods of statistics (Princeton University Press, Princeton, N. J., 1946).

WASHINGTON, February 8, 1954.

4. Functional Relationships

Foreword

In 1959, Forman S. Acton wrote a book of 267 pages on the Analysis of Straight-line Data. In his preface he admitted that there are "... important problems for which, unfortunately, no adequate answers have been found." Ten years later we have found new problems added to the unsolved old problems. Apparently the subject of relationship between variables, even for the seemingly simple straight line case, is far from being exhausted.

Paper (4.1) in this section is an example taken from John Mandel's doctoral dissertation, and illustrates how the same set of data can be scrutinized on the basis of eight different assumed models.

Chapter 5 of NBS Handbook 91 presents a general picture (4.2) of the physical situations which can be described by a linear relationship between two variables, and gives the uses and interpretations of the resulting equation fitted under the different models assumed. Its table 5.1 is a summary of the cases that are usually encountered by experimenters involved in physical measurements.

Mandel's other paper (4.3) on calibration curves points out the difficulty resulting "from the interdependence of multiple conclusions drawn from the same data, especially when there is a strong correlation between the parameters involved." He has treated this subject in further detail in Statistical Analysis of Experimental Data (Selected References B6).

There are few papers available that deal with polynomial or other types of curve fitting. Hockersmith and Ku (4.4) demonstrate the use of a quadratic curve in interpreting data on proving rings. For further reading, Chapter 6 of Handbook 91 (Selected Reference C2) and Draper and Smith (Selected Reference B8) are recommended.

With the availability of canned computer programs, "least squares" has become a magic term. Eisenhart's paper (4.5) gives a historical account of the evolution of the meaning of "least" from the days of Laplace and Gauss, and is interesting and pertinent reading.

One word of warning! Be sure you use a computer program which has been adequately tested for round-off errors when fitting a polynomial (or multi-variable) equation to a set of data. Some popular programs using naive matrix methods have been found to yield accuracies of only one significant digit in the coefficients of a 5th degree equation fitted to 21 equally spaced points, whereas more sophisticated routines would produce five significant digits.

A Statistical Study of

Physical Classroom Experiments

First Example: The Acceleration of Gravity, g

John Mandel

1. Principle and Method of Measurement

A pendulum is constructed by suspending a metal sphere of about 3 cm. in diameter by means of a thread of negligible weight and elongation. The length of the pendulum, from the point of suspension of the thread to the center of gravity of the sphere, is determined by means of a measuring tape. The pendulum is made to swing with an amplitude not exceeding 10 percent of its length, and the time for 50 oscillations is recorded three times in succession.

The measurements are carried out for 5 values of l (the length of the pendulum), equal approximately to 175, 150, 125, 100 and 75 cm. The students are instructed to plot T^2 versus l, T being the period of the oscillation, and determine g, the acceleration of gravity, by using the relation

$$T = 2\pi\sqrt{\frac{l}{g}} . \qquad (1a)$$

Thus, g is calculated from the slope, equal to $\frac{4\pi^2}{g}$, of the straight line relating T^2 to l.

2. The Data

Table 1 lists for each of 10 students, the five values of l with the corresponding measured values for 50T. Each value of 50T is actually the average of 3 replicate determinations, using the same value of l. Student no.6 made no measurements for l = 150 cm.

Table 1

Basic Data for the Determination of g

Student		Measurements (1)				
1	ℓ	175.2	151.5	126.4	101.7	77.0
	50 T	132.5	123.4	112.8	101.2	88.2
2	ℓ	179.0	150.0	125.0	100.0	75.0
	50 T	133.7	122.3	111.3	99.8	85.8
3	ℓ	170.0	149.3	124.8	100.4	76.4
	50 T	130.8	122.5	112.1	100.5	87.6
4	ℓ	165.1	149.8	125.0	100.0	75.0
	50 T	129.0	122.8	112.2	100.0	86.8
5	ℓ	171.5	150.0	124.9	100.0	75.0
	50 T	131.1	122.6	111.9	100.1	86.8
6	ℓ	175.8	-	125.0	99.7	75.0
	50 T	132.8	-	112.0	100.1	86.8
7	ℓ	172.0	150.0	125.0	100.0	75.0
	50 T	131.6	122.8	112.0	100.2	86.8
8	ℓ	175.3	149.9	125.0	100.0	75.0
	50 T	132.1	122.2	111.8	100.0	86.5
9	ℓ	165.5	150.0	125.0	100.0	75.0
	50 T	128.9	122.7	112.0	100.2	86.9
10	ℓ	175.8	150.7	125.8	100.8	74.6
	50 T	132.8	123.0	112.3	100.7	86.7

(1) ℓ is expressed in cm.; 50 T in seconds. Each value for 50 T is the average of triplicate determinations; the standard deviation among triplicates is 0.080.

3. Analysis of the Data: Part I

For each student, the value of T^2 was calculated for each of his ℓ -values. Using the method of least squares for linear regression, which is described in most textbooks of statistics (see for example [4,10]), a straight line was fitted to the (T^2,ℓ) points, as follows

$$T^2 = \alpha + \beta\ell. \qquad (1b)$$

The quantity α is the intercept of the line, β its slope. This equation is slightly more general than Eq. (1a), which it contains as a special case, namely when $\alpha = 0$. The reasons for using this procedure will be explained below.

The results are summarized in Table 2, which lists the intercepts, slopes, and residual standard deviations for all regression lines. Also listed are, for reasons to be discussed later, the ordinates of the fitted lines for ℓ = 125 cm.

At this point an excellent opportunity arises for the students to attempt to formulate questions that are pertinent in terms of the physical theory (Equation (1)), underlying the experiment. The instructor can then show how these questions are translated in statistical terminology and explain the statistical methodology used for their elucidation. We will illustrate this point by posing the following questions :

(1) Are there systematic differences between the regression lines for the different students ?

(2) If such differences are found, are they due uniquely to differences between the intercepts, or are the slopes also different ?

(3) How do the slopes compare with the "theoretical" value, $\frac{4\pi^2}{g}$, where g is given its known value for the Netherlands, $g = 981.3 \frac{cm}{sec^2}$, thus making the theoretical slope equal to

Table 2

Results for Regression of T^2 on ℓ

Student	Number of points	Intercept (sec^2)	Slope (sec^2/cm)	Standard Deviation of Residuals (sec^2)	Height[1] (sec^2)
1	5	0.045	3.9863×10^{-2}	0.0067	5.0280
2	5	- 0.072	4.0345	0.0165	4.9711
3	5	- 0.006	4.0278	0.0050	5.0290
4	5	- 0.033	4.0500	0.0120	5.0294
5	5	0.009	4.0024	0.0029	5.0124
6	4	0.011	4.0065	0.0024	5.0167
7	5	- 0.016	4.0333	0.0064	5.0256
8	5	0.024	3.9708	0.0099	4.9876
9	5	0.013	4.0065	0.0035	5.0207
10	5	0.022	3.9992	0.0057	5.0210

(1) Ordinate of fitted line for ℓ = 125 cm.

4.0231×10^{-2} (in $\frac{sec^2}{cm}$) ?

(4) How do the results compare with the theoretical straight line which, in addition to having a slope of 4.0231×10^{-2} sec^2/cm, must also have a zero-intercept ?

The students can be shown at this point that the general statistical theory for fitting linear models provides a powerful and elegant tool for answering these questions [1]. The basic idea underlying this procedure is to "embed" the model that is to be tested into a more general linear model, i.e. one with a larger number of estimated parameters. An elementary exposition may be found in [10].

Denoting the length by x and the square of the period, T^2, by y, Equation (1a) can be written

$$E(y) = \frac{4\pi^2}{g} x , \qquad (2)$$

where E(y) represents the "expected value" of y, i.e. the value of the y freed of random experimental error.

If g is given its theoretical value, $g = 981.3$, model (2) involves no unknown parameters and becomes

$$E(y) = (4.0231 \times 10^{-2}) x . \qquad (3)$$

This model can be embedded in the slightly more general model

$$E(y) = \beta x , \qquad (4)$$

---

(1) It is assumed here that the students have a sufficient background in statistical theory to follow such an analysis. For students with a lesser background, the instructor can proceed at once with the control chart analysis (section 4).

Schematic Representation of the Statistical Testing Process

Symbol	Model	Number of Parameters	Residual Degrees of Freedom	Sum of Squares of Residuals (10^6 sec^2)	Mean Square of Residuals (10^6 sec^2)
(A)	$E(y) = \alpha_i + \beta_i x$	20	29	2042	70.4
(B)	$E(y) = \alpha + \beta_i x$	11	38	5763	152
(B')	$E(y) = \beta_i x$	10	39	5768	148
(C)	$E(y) = \alpha_i + \beta x$	11	38	5207	137
(C')	$E(y) = \alpha_i + (4.0231\ 10^{-2})x$	10	39	6037	155
(D)	$E(y) = \alpha + \beta x$	2	47	22169	472
(G)	$E(y) = \beta x$	1	48	22172	462
(H)	$E(y) = (4.0231\ 10^{-2})x$	0	49	33476	683

Table 3

where β may differ from the theoretical value 4.0231×10^{-2}. Model (4) can in turn be embedded in one allowing for a non-zero intercept

$$E(y) = \alpha + \beta x. \tag{5}$$

It is conceivable that either α, or β, or both, vary from student to student, in which case we obtain the models

$$\begin{cases} E(y) = \alpha_i + \beta x & (6a) \\ E(y) = \alpha + \beta_i x, & (6b) \end{cases}$$

(where the subscript i refers to the ith student) or the more general model

$$E(y) = \alpha_i + \beta_i x. \tag{7}$$

Equation (6a) contains the interesting sub-model

$$E(y) = \alpha_i + (4.0231 \times 10^{-2})\, x \tag{8}$$

and Equation (6b) similarly includes the case

$$E(y) = 0 + \beta_i x. \tag{9}$$

Now, while a physicist would probably start by assuming that Equation (3) holds, and only abandon this hypothesis if it is definitely contradicted by the data, the statistician would generally choose the inverse path. In other words, the statistician would start with the most general (and therefore safest) assumption expressed by Equation (7), and then attempt to particularize it gradually, i.e. reduce gradually the number of parameters to be estimated, using the data as a criterion for the validity of each step in the reduction process. The process is schematically re-

Table 4

Basic Calculations

Student	X	Y	U	W	P
1	631.8	25.411	85,896.14	138.776639	3452.5794
2	629.0	25.017	85,791.00	136.016095	3415.9500
3	620.9	24.980	82,682.65	133.851690	3326.7410
4	614.9	24.738	80,948.05	131.132852	3258.0492
5	621.4	24.918	83,137.26	133.648110	3333.3366
6	475.5	19.094	62,095.73	100.087500	2492.9908
7	622.0	25.007	83,334.00	134.761129	3351.1440
8	625.2	24.946	84,450.10	134.355178	3368.4217
9	615.5	24.723	81,140.25	130.868821	3258.6380
10	627.7	25.213	85,167.57	137.320741	3419.8318
Sum	6083.9	244.047	814,642.75	1310.818755	32,677.6825

Student	u	w	p
1	6061.89	9.632855	241.6454
2	6662.80	10.846037	268.8114
3	5579.29	9.051610	224.7246
4	5327.65	8.739124	215.7700
5	5909.67	9.466765	236.5276
6	5570.67	8.942291	223.1916
7	5957.20	9.691119	240.2732
8	6275.09	9.894595	249.1739
9	5372.20	8.623475	215.2367
10	6366.11	10.181667	254.5918
Sum	59,082.57	95.069538	2369.9462

presented in Table 3. The statistical analysis consists in testing each of the successive models against one in which it can be embedded, starting from the top and proceeding gradually downward.

The basic numerical material consists of the quantities

$$\Sigma x \quad \text{and} \quad \Sigma x^2, \quad \Sigma y \quad \text{and} \quad \Sigma y^2, \quad \Sigma xy,$$

which we denote respectively by the symbols

$$X = \Sigma x, \quad U = \Sigma x^2, \quad Y = \Sigma y, \quad W = \Sigma y^2, \quad P = \Sigma xy \tag{10}$$

and of the derived quantities

$$\left.\begin{aligned} u &= \Sigma (x - \bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = U - \frac{X^2}{n} \\ w &= \Sigma (y - \bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = W - \frac{Y^2}{n} \\ p &= \Sigma (x - \bar{x})(y - \bar{y}) = \Sigma xy - \frac{\Sigma x \, \Sigma y}{n} = P - \frac{XY}{n} . \end{aligned}\right\} \tag{11}$$

where n is the number of experimental points in a regression line. These quantities are tabulated in Table 4 for each of the ten students.

Table 3 also gives the degrees of freedom, sums of squares and mean squares of the residuals for all of the models considered (*). Using the latter as a guide one readily finds the model into which each model to be tested is to be embedded. Thus it is clear that (B) and (C) are both tested against (A); (B') is tested against (B), and (C') against (C). The model (D) can be embedded in (A), (B), or (C); which of these is chosen will depend on the outcome of the tests for models (B) and (C). Model (G) can be embedded in (D), and model (H) in (G).

(*) The only model for which the computations are not directly apparent is (B). The appropriate formulas for this model are given in Appendix A.

In practice it is unlikely that many of these tests will have to be made, since any finding of significance will generally make subsequent reduction steps of academic interest only. Thus, in the case of our data, models (B) and (C) are both unacceptable and it is therefore unnecessary to continue the statistical testing process to the more specialized models. In testing any hypothesis, such as for example (C), one first calculates the reduction in the sum of squares from (C) to the more general model (A), divides this by the corresponding reduction in the degrees of freedom and compares this mean square, by means of the F test, to that corresponding to the more general model (A). Thus, for testing (C) we have

$$F = \frac{(5207 - 2042)/(38 - 29)}{70.4} = \frac{352}{70.4} = 5.00$$

with 9 and 29 degrees of freedom.

The conclusion of the statistical analysis is that the data are not consistent with the hypothesis that all the students obtained the theoretical relation between T^2 and $\underline{\ell}$, nor even with the hypothesis that they all obtained the same (incorrect) relationship. It is also seen that both the slopes and the intercepts vary from student to student. The only acceptable model is (A), which associates with each student an individual relationship between T^2 and $\underline{\ell}$.

It is indispensible, before accepting this hypothesis, to examine the residuals individually, and to verify that they do not display striking patterns of non-randomness. Table 5, which lists the residuals, throws no serious suspicion on the validity of model (A).

After performing the analysis based on the general linear hypothesis, it is well to point out that while this method is elegant and powerful, it fails to provide detailed information about the results of each individual student. An excellent way of obtaining the latter consists in carrying out a control chart type of analysis.

Table 5

Residuals from Model (A) (1)

Student

Approx. ℓ	1	2	3	4	5	6	7	8	9	10
175	-6	0	2	2	2	0	6	-5	2	2
150	7	3	-6	-2	-1	-	-2	-3	-1	3
125	5	-16	5	7	1	-1	-8	12	-3	- 9
100	-2	22	2	-17	-3	3	-1	5	-3	3
75	-2	-9	-3	10	3	-2	5	-9	3	2

(1) All residuals were multiplied by 10^3.

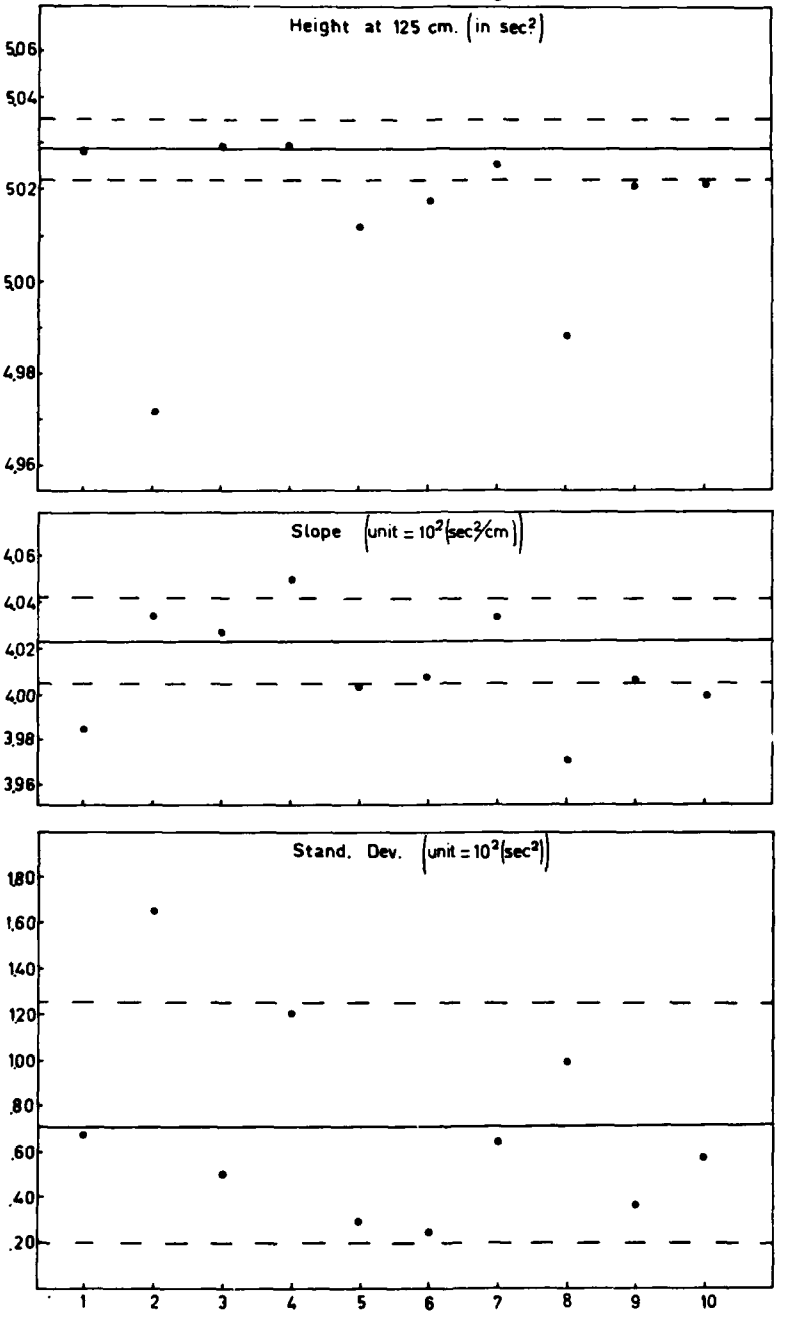
4. Analysis of the Data; Part II

Table 2 is the starting point of the second part of the analysis. It may be observed that the situation is quite analogous to that encountered in the evaluation of interlaboratory test results [10,11]. The central idea in such an evaluation is the setting up of a model containing parameters that may vary from laboratory to laboratory (in the present case : from student to student), and to display the variation of the parameters by means of control charts [1]. Table 2, when viewed as a representation of model (A), contains three variable parameters : the intercept, the slope, and the standard deviation about the regression line. However, the intercept and the slope are highly correlated, in a statistical sense. This means that the types of information provided by these two parameters largely overlap. For this reason it is advisable to replace one of these parameters, for example the intercept, by one that is independent of the other (the slope). In our case, the ordinate of the regression line at ℓ = 125 cm is, for all students except student no.6, very close to the ordinate of the centroid of the line, and therefore essentially independent of the slope. For these reasons, the control charts shown in Figure 1 are those of the ordinate at ℓ = 125 cm (which we shall call the "height" of the line), the slope, and the standard deviation about the regression line.

The control lines, which are all "two-sigma lines", are based, for all three charts, on the average of the standard deviations for T^2: $\bar{\sigma}$ = 0.0071 sec^2. The central lines for the height and the slope were taken each at their theoretical value (i.e. the values required by Eq. (1), assuming g = 981.3).
Control lines for standard deviations may be derived from the chi-square distribution, using the relation $ns^2/\sigma^2 = \chi^2_n$, where n is the number of degrees of freedom and χ^2_n the chi-square variate with n degrees of freedom. From this relation it follows that 95 percent control lines may be calculated from the double inequality

$$\sigma\left(\frac{\chi_a}{\sqrt{n}}\right) < s < \sigma\left(\frac{\chi_b}{\sqrt{n}}\right)$$

Fig. 1: Control Chart Analysis for Determiniation of g
Centroid, slope and std. dev. for regression lines of T^2 on l.



where σ may be approximated by the average value of the standard deviations and χ_a and χ_b are respectively the 2.5 and the 97.5 percentiles of chi. These percentile-values may be found in the Biometrika tables [3]. The number of degrees of freedom in our case is 3, except for student 6. The calculations were made, using n = 3. The standard errors for the height and the slope were computed using the classical formulas ($\sigma/\sqrt{n}$ and $\sigma\sqrt{\frac{1}{\Sigma (x-\bar{x})^2}}$), in which an average value was taken for $\Sigma(x-\bar{x})^2$ (omitting student 6 in the average).

The general picture emerging from these charts is one of considerable skepticism about the work of the students in this test. The standard deviation for one student is suspect. For the height, only four students show values that are not significantly different from theory, and for the slope only five students agree with theory to within the residual error. For only two students do the values for both parameters agree with theory.

The important question is of course to discover the physical causes (shortcomings in the experiment) that led to this state of affairs. The students should be encouraged to offer suggestions.

It is also interesting to compare the average standard deviation, $\bar{\sigma}$ = 0.0071 sec, with that expected from the known experimental errors in measuring $\underline{\ell}$ and T. The estimate $\bar{\sigma}$ is that obtained from a regression of T^2 on $\underline{\ell}$. The replication standard deviation for 50 T was found to be 0.080 sec, for an average (of 50 T) of about 112 sec. Thus the coefficient of variation (C.V.) is 0.080/112 = $= 7.1 \times 10^{-4}$. Therefore, the C.V. for T^2 is 14.2×10^{-4}. The average of T^2 being about 5 sec^2, we therefore have an expected $\sigma_{T^2} = 5 \times 14.2 \times 10^{-4} = 71 \times 10^{-4}$ sec^2. Since averages of triplicates were used, the standard error of each plotted point (assuming $\underline{\ell}$ to be free of error) is $\frac{71 \times 10^{-4}}{\sqrt{3}} = 0.0041$ sec^2. Actually, $\underline{\ell}$ was not free of error, and its error is reflected in the scatter about the regression line. Assuming an uncertainly range of about 1.0 cm for $\underline{\ell}$ (the students reported $\pm$ 0.2 to $\pm$ 0.5 cm), the

standard deviation would be of the order of $\frac{1.0}{5}$ = 0.2 cm. Since the slope of T^2 versus $\underline{\ell}$ is about 0.04 sec^2/cm, this corresponds to a standard deviation along the ordinate of 0.04 x 0.2 = 0.008. Thus, the total standard deviation expected about the regression line is about $\sqrt{(0.0041)^2 + (0.0080)^2}$ = 0.0090 sec^2, which is comparable to the observed average $\bar{\sigma}$ = 0.0071 sec^2. Thus, the observed scatter about the regression lines is consistent with the estimated error, confirming once more the adequacy of model (A).

In conclusion, the analysis indicates the presence of unexplained systematic errors for almost all students.

5. Critique of the Experimental Design

This experiment presents an opportunity to raise the general question of the relationship between design and analysis. While the target values for $\underline{\ell}$ are specifically given in the instruction manual used by the students, nothing is said about the desired closeness of the actual values selected by the student to those values.

Two rather different experimental situations can arise:

1) the values of $\underline{\ell}$ are fixed, and the student is instructed to approach them as accurately as he can; or
2) $\underline{\ell}$ is set only roughly near the target values, but measured as accurately as possible.

In the first design, a regression analysis of T^2 versus $\underline{\ell}$, using the ordinary equations (x-variable free of error), is justified, as was shown by Berkson [2,10]. In the second design, both T^2 and $\underline{\ell}$ are subject to error (though their errors are uncorrelated). In that case, the regression calculations are somewhat more complex, and much of the simplicity and elegance of the statistical analysis

is lost, unless the error of $\underline{\ell}$ is made to be negligible in comparison with that of T^2.

Suppose that, in accordance with the instruction manual, g is determined from the slope of the regression of T^2 on $\underline{\ell}$, and that $\underline{\ell}$ is either a controlled variable, in the Berksonian sense, or is measured with negligible error. Then, in order to avoid the complications of a weighted regression analysis, the variance of T^2 should be the same for all $\underline{\ell}$. Let $\underline{t}$ be the time required for $\underline{n}$ complete oscillation of the pendulum. Then we have

$$t = n\,T, \tag{12}$$

$$\sigma_t = n\,\sigma_T . \tag{13}$$

We require that

$$\sigma_{T^2} = \text{constant}, \tag{14}$$

hence $\sigma_{T^2} = 2T\,\sigma_T = \text{constant}.$ (15)

Introducing Eq.(13) in Eq.(14) we obtain

$$\sigma_{T^2} = 2T\sigma_T = 2T\,\frac{\sigma_t}{n} = \text{constant}. \tag{16}$$

We may assume that the standard deviation of the time-measurement, σ_t, is a constant. Then, Eq.(16) requires that $\underline{n}$ be taken proportional to T. This is equivalent to requiring that $\underline{t}$ be proportional to T^2, i.e. to $\underline{\ell}$. This result may seem surprising, inasmuch as, for a determination of g in accordance with Eq.(1), the relative error $\underline{\ell}$ is already more disturbing for small $\underline{\ell}$, and this is now aggravated by making the relative error for the time measurement also larger for small $\underline{\ell}$. The answer to this apparent paradox is of course that in the present procedure, g is not determined directly from Eq.(1), but rather from the slope of a regression line of T^2 on $\underline{\ell}$. Th nstructor can use this opportunity to further

stress the important relationship that always exists between the design of an experiment and the manner in which the data will be analyzed.

The preceding discussion shows that the number of oscillations for which the total time was measured should have been different for the different lengths of the pendulum: they should have been taken proportunally to $\sqrt{\ell}$, in order to make the unweighted regression analysis strictly valid.

Finally, it should be noted that the usual precautions of randomization, to avoid systematic errors, were not observed in the experiment here described. Thus, the five values of ℓ should not have been taken consistently in the order 175, 150, 125, 100 and 75 cm. The student should be shown that to do so may introduce fictitious changes in the slope and the intercept of the regression line, due to possible trends in the measuring technique.

Reprinted from: A Statistical Study of Physical Classroom Experiments, Technische Hogeschool Eindhoven, 17-33, 1965

CHAPTER 5 *

CHARACTERIZING LINEAR RELATIONSHIPS BETWEEN TWO VARIABLES

Mary G. Natrella

5-1 INTRODUCTION

In many situations it is desirable to know something about the relationships between two characteristics of a material, product, or process. In some cases, it may be known from theoretical considerations that two properties are functionally related, and the problem is to find out more about the structure of this relationship. In other cases, there is interest in investigating whether there exists a degree of association between two properties which could be used to advantage. For example, in specifying methods of test for a material, there may be two tests available, both of which reflect performance, but one of which is cheaper, simpler, or quicker to run. If a high degree of association exists between the two tests, we might wish to run regularly only the simpler test.

In this chapter, we deal only with linear relationships. Curvilinear relationships are discussed in Chapter 6 (see Paragraph 6-5). It is worth noting that many nonlinear relationships may be expressed in linear form by a suitable transformation (change of variable). For example, if the relationship is of the form $Y = aX^b$, then $\log Y = \log a + b \log X$. Putting $Y_T = \log Y$, $b_0 = \log a$, $b_1 = b$, $X_T = \log X$, we have the linear expression $Y_T = b_0 + b_1 X_T$ in terms of the new (transformed) variables X_T and Y_T.

A number of common linearizing transformations are summarized in Table 5-4 and are discussed in Paragraph 5-4.4.

5-2 PLOTTING THE DATA

Where only two characteristics are involved, the natural first step in handling the experimental results is to plot the points on graph paper. Conventionally, the *independent variable* X is plotted on the horizontal scale, and the *dependent variable* Y is plotted on the vertical scale.

There is no substitute for a plot of the data to give some idea of the general spread and shape of the results. A pictorial indication of the probable form and sharpness of the relationship, if any, is indispensable and sometimes may save needless computing. When investigating a structural relationship, the plotted data will show whether a hypothetical linear relationship is borne out; if not, we must consider whether there is any theoretical basis for fitting a curve of higher degree. When looking for an empirical association of two characteristics, a glance at the plot will reveal whether such association is likely or whether there is only a patternless scatter of points.

In some cases, a plot will reveal unsuspected difficulties in the experimental setup which must be ironed out before fitting any kind of relationship. An example of this occurred in

* NBS Handbook 91, 1966.

measuring the time required for a drop of dye to travel between marked distances along a water channel. The channel was marked with distance markers spaced at equal distances, and an observer recorded the time at which the dye passed each marker. The device used for recording time consisted of two clocks hooked up so that when one was stopped, the other started: Clock 1 recorded the times for Distance Markers 1, 3, 5, etc.; and Clock 2 recorded times for the even-numbered distance markers. When the elapsed times were plotted, they looked somewhat as shown in Figure 5-1. It is obvious that there was a systematic time difference between odd and even markers (presumably a lag in the circuit connecting the two clocks). One could easily have fitted a straight line to the odd-numbered distances and a different line to the even-numbered distances, with approximately constant difference between the two lines. The effect was so consistent, however, that the experimenter quite properly decided to find a better means of recording travel times before fitting any line at all.

If no obvious difficulties are revealed by the plot, and the relationship appears to be linear, then a line $Y = b_0 + b_1X$ ordinarily should be fitted to the data, according to the procedures given in this Chapter. Fitting by eye usually is inadequate for the following reasons:

(a) No two people would fit exactly the same line, and, therefore, the procedure is not objective;

(b) We always need some measure of how well the line does fit the data, and of the uncertainties inherent in the fitted line as a representation of the true underlying relationship—and these can be obtained only when a formal, well-defined mathematical procedure of fitting is employed.

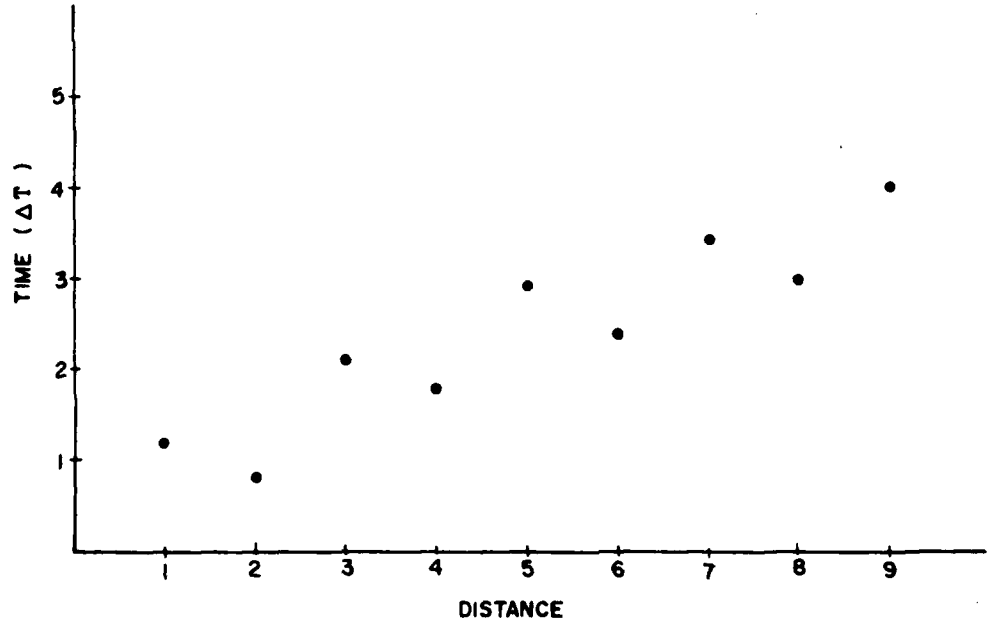


Figure 5-1. Time required for a drop of dye to travel between distance markers.

5-3 TWO IMPORTANT SYSTEMS OF LINEAR RELATIONSHIPS

Before giving the detailed procedure for fitting a straight line, we discuss different physical situations which can be described by a linear relationship between two variables. The methods of description and prediction may be different, depending upon the underlying system. In general, we recognize two different and important systems which we call *Statistical* and *Functional*. It is not possible to decide which is the appropriate system from looking at the data. The distinction must be made before fitting the line—indeed, before taking the measurements.

5-3.1 FUNCTIONAL RELATIONSHIPS

In the case of a Functional Relationship, there exists an exact mathematical formula (y as a function of x) relating the two variables, and the only reason that the observations do not fit this equation exactly is because of disturbances or errors of measurement in the observed values of one or both variables. We discuss two cases of this type:

FI—Errors of measurement affect only one variable (Y). (See Fig. 5-2).

FII—Both variables (X and Y) are subject to errors of measurement. (See Fig. 5-3).

Common situations that may be described by Functional Relationships include calibration lines, comparisons of analytical procedures, and relationships in which time is the X variable.

For instance, we may regard Figure 5-2 as portraying the calibration of a straight-faced spring balance in terms of a series of weights whose masses are accurately known. By Hooke's Law, the extension of the spring, and hence the position y of the scale pointer, should be determined exactly by the mass x upon the pan through a linear functional relationship* $y = \beta_0 + \beta_1 x$. In practice, however, if a weight of mass x_1 is placed upon the pan repeatedly and the position of the pointer is read in each instance, it usually is found that the readings Y_1 are not identical, due to variations in the performance of the spring and to reading errors. Thus, corresponding to the mass x_1 there is a distribution of pointer readings Y_1; corresponding to mass x_2, a distribution of pointer readings Y_2; and so forth—as indicated in Figure 5-2. It is customary to assume that these distributions are normal (or, at least symmetrical and all of the same form) and that the mean of the distribution of Y_i's coincides with the *true value* $y_i = \beta_0 + \beta_1 x_i$.

If, instead of calibrating the spring balance in terms of a series of accurately known weights, we were to calibrate it in terms of another spring balance by recording the corresponding pointer positions when a series of weights are placed first on the pan of one balance and then on the pan of the other, the resulting readings (X and Y) would be related by a linear structural relationship FII, as shown in Figure 5-3, inasmuch as both X and Y are affected by errors of measurement. In this case, corresponding to the repeated weighings of a single weight w_1 (whose true mass need not be known), there is a joint distribution of the pointer readings (X_1 and Y_1) on the two balances, represented by the little transparent *mountain* centered over the *true point* (x_1, y_1) in Figure 5-3; similarly at points (x_2, y_2) and (x_3, y_3), corresponding to repeated weighings of other weights w_2 and w_3, respectively. Finally, it should be noticed that this FII model is more general than the FI model in that it does *not* require linearity of response of each instrument to the independent variable w, but merely that the response curves

* *Note on Notation for Functional Relationships:*
We have used x and y to denote the true or accurately known values of the variables, and X and Y to denote their values measured with error. In the FI Relationship, the independent variable is always without error, and therefore in our *discussions* of the FI case and in the paragraph headings we always use x. In the Worksheet, and Procedures and Examples for the FI case, however, we use X and Y because of the computational similarity to other cases discussed in this Chapter (i.e., the computations for the Statistical Relationships).
In the FII case, both variables are subject to error, and clearly we use X and Y everywhere for the observed values.

of the two instruments be linearly related, that is, that $X = a + b \cdot f(w)$ and $Y = c + d \cdot f(w)$, where $f(w)$ may be linear, quadratic, exponential, logarithmic, or whatever.

Table 5-1 provides a concise characterization of FI and FII relationships. Detailed problems and procedures with numerical examples for FI relationships are given in Paragraphs 5-4.1 and 5-4.2, and for FII relationships in Paragraph 5-4.3.

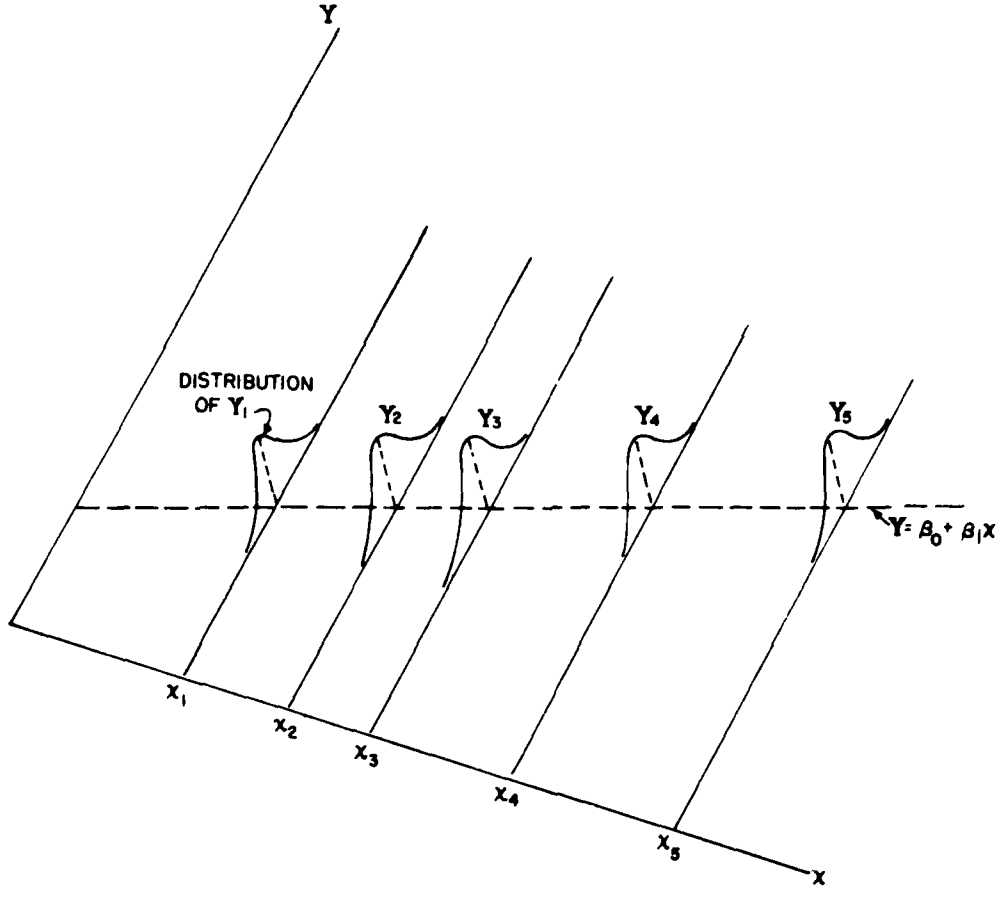


Figure 5-2. Linear functional relationship of Type FI (only Y affected by measurement errors).

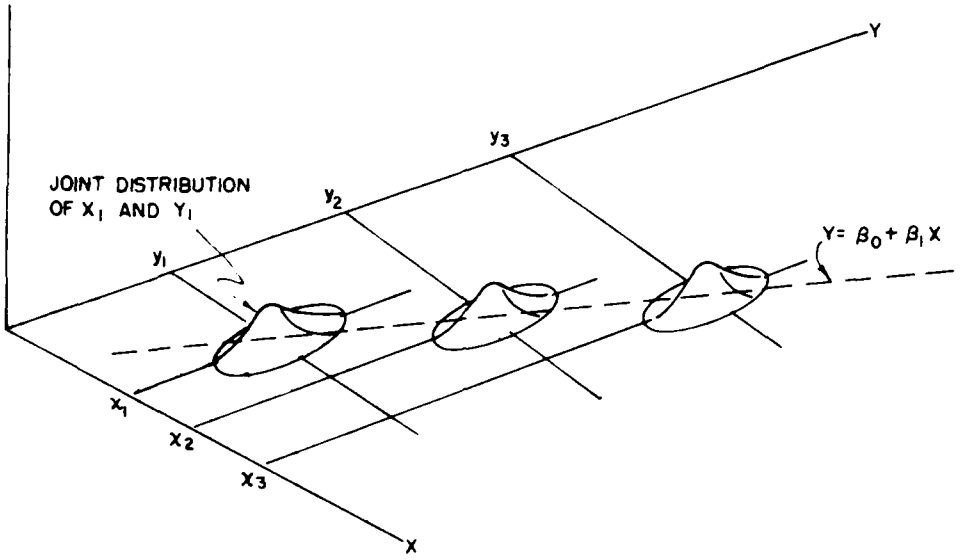


Figure 5-3. Linear functional relationship of Type FII (both X and Y affected by measurement errors).

5-3.2 STATISTICAL RELATIONSHIPS

In the case of a Statistical Relationship, there is no exact mathematical relationship between X and Y; there is only a statistical association between the two variables as characteristics of individual items from some particular population. If this statistical association is of bivariate normal type as shown in Figure 5-4, then the *average* value of the Y's associated with a particular value of X, say $\bar{Y}_X$, is found to depend linearly on X, i.e., $\bar{Y}_X = \beta_0 + \beta_1 X$; similarly, the *average* value of the X's associated with a particular value of Y, say $\bar{X}_Y$, depends linearly on Y (Fig. 5-4) i.e., $\bar{X}_Y = \beta_0' + \beta_1' Y$; but—and this is important!—the two lines are *not* the same, i.e., $\beta_1' \neq \frac{1}{\beta_1}$ and $\beta_0' \neq -\frac{\beta_0}{\beta_1}$.*

* Strictly, we should write

$m_{Y \cdot X} = \beta_0 + \beta_1 X$,

and

$m_{X \cdot Y} = \beta_0' + \beta_1' Y$

to conform to our notation of using m to signify a population mean. But this more exact notation tends to conceal the parallelism of the curve-fitting processes in the FI and SI situations. Consequently, to preserve appearances here and in the sequel, we use $\bar{Y}_X$ in place of $m_{Y \cdot X}$ and $\bar{X}_Y$ in place of $m_{X \cdot Y}$—and it should be remembered that these signify *population means*.

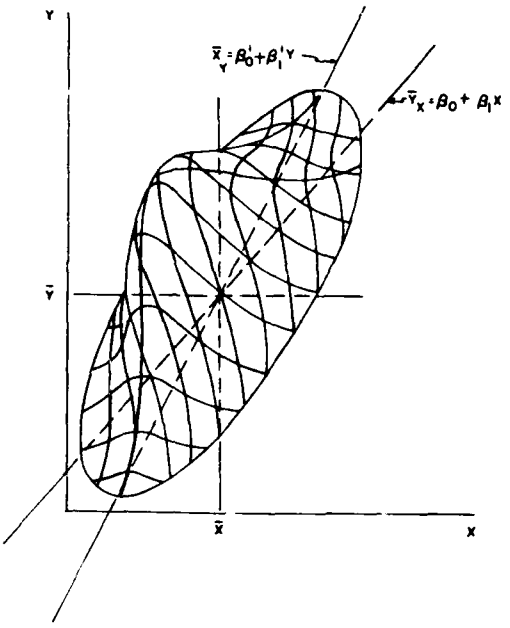


Figure 5-4. A normal bivariate frequency surface.

If a random sample of items is drawn from the population, and the two characteristics X and Y are measured on each item, then typically it is found that errors of measurement are negligible in comparison with the variation of each characteristic over the individual items. This general case is designated SI. A special case (involving preselection or restriction of the range of one of the variables) is denoted by SII.

SI Relationships. In this case, a random sample of items is drawn from some definite population (material, product, process, or people), and two characteristics are measured on each item.

A classic example of this type is the relationship between height and weight of men. Any observant person knows that weight tends to vary with height, but also that individuals of the same height may vary widely in weight. It is obvious that the errors made in measuring height or weight are very small compared to this inherent variation between individuals. We surely would not expect to predict the exact weight of one individual from his height, but we might expect to be able to estimate the average weight of all individuals of a given height.

The height-weight example is given as one which is universally familiar. Such examples also exist in the physical and engineering sciences, particularly in cases involving the interrelation of two test methods. In many cases there may be two tests that, strictly speaking, measure two basically different properties of a material, product, or process, but these properties are statistically related to each other in some complicated way and both are related to some performance characteristic of particular interest, one usually more directly than the other. Their interrelationship may be obscured by inherent variations among sample units (due to varying density, for example). We would be very interested in knowing whether the relationship between the two is sufficient to enable us to predict with reasonable accuracy, from a value given by one test, the average value to be expected for the other—particularly if one test is considerably simpler or cheaper than the other.

The choice of which variable to call X and which variable to call Y is arbitrary—actually there are two regression lines. If a statistical association is found, ordinarily the variable which is easier to measure is called X. Note well that this is the only case of linear relationship in which it may be appropriate to fit two different lines, one for predicting Y from X and a different one for predicting X from Y, *and* the only case in which the sample correlation coefficient r is meaningful as an estimate of the degree of association of X and Y *in the population* as measured by the population coefficient of correlation $\rho = \sqrt{\beta_1 \beta_1'}$. The six sets of contour ellipses shown in Figure 5-5 indicate the manner in which the location, shape, and orientation of the normal bivariate distribution varies with changes of the population means (m_X and m_Y) and standard deviations (σ_X and σ_Y) of X and Y and their coefficient of correlation in the population (ρ_{XY}).

If $\rho = \pm 1$, all the points lie on a line and $Y = \beta_0 + \beta_1 X$ and $X = \beta_0' + \beta_1' Y$ coincide. If $\rho = +1$, the slope is positive, and if $\rho = -1$, the slope is negative. If $\rho = 0$, then X and Y are said to be uncorrelated.

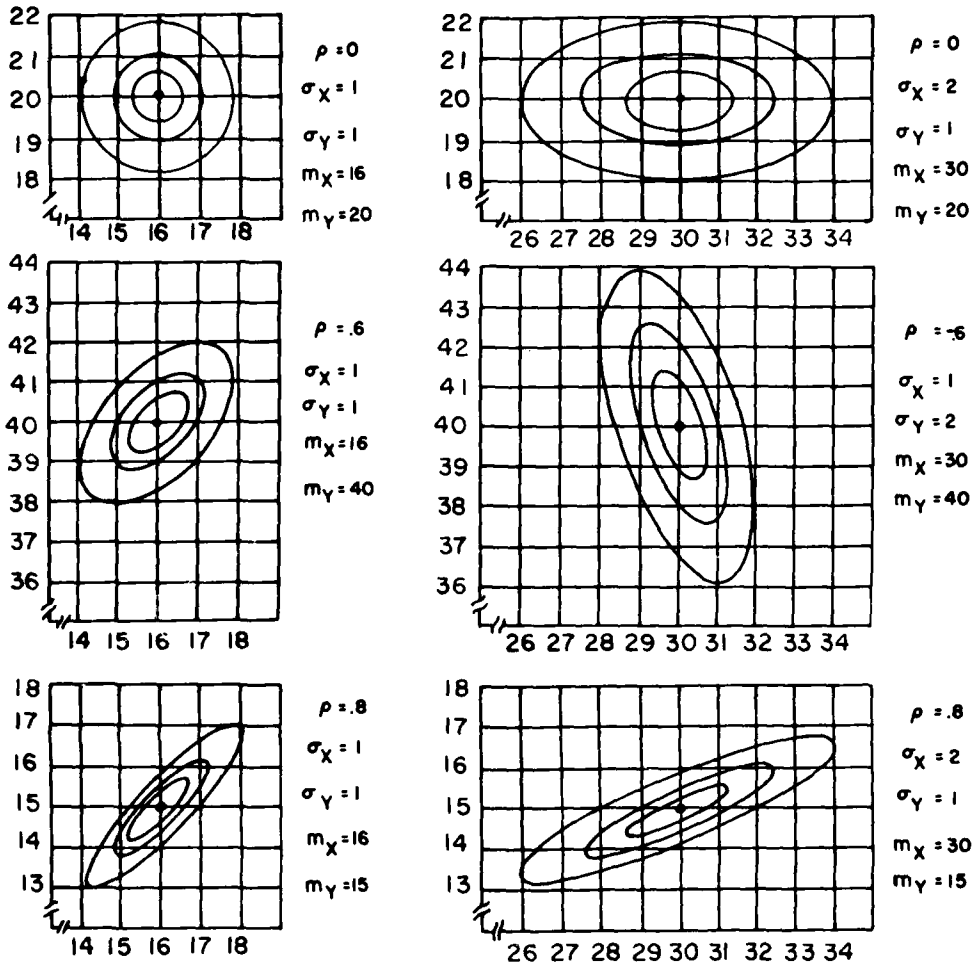


Figure 5-5. Contour ellipses for normal bivariate distributions having different values of the five parameters m_X, m_Y, σ_X, σ_Y, ρ_{XY}.

Adapted with permission from *Statistical Inference* by Helen M. Walker and Joseph Lev, copyright, 1953, Holt, Rinehart and Winston, Inc., New York, N. Y.

SII Relationships. The general case described above (SI) is the most familiar example of a statistical relationship, but we also need to consider a common case of Statistical Relationship (SII) that must be treated a bit differently. In SII, one of the two variables, although a random variable in the population, is sampled only within a limited range (or at selected preassigned values). In the height-weight example, suppose that the group of men included only those whose heights were between 5'4" and 5'8". We now are able to fit a line predicting weight from height, but are unable to determine the correct line for predicting height from weight. A correlation coefficient computed from such data is not a measure of the true correlation among height and weight in the (unrestricted) population.

The restriction of the range of X, when it is considered as the independent variable, does

not spoil the estimates of $\bar{Y}_X$ when we fit the line $\bar{Y}_X = b_0 + b_1X$. The restriction of the range of the dependent variable (i.e., of Y in fitting the foregoing line, or of X in fitting the line $\bar{X}_Y = b_0' + b_1'Y$), however, gives a seriously distorted estimate of the true relationship. This is evident from Figure 5-6, in which the contour ellipses of the top diagram serve to represent the bivariate distribution of X and Y in the unrestricted population, and the "true" regression lines of $\bar{Y}_X$ on X and $\bar{X}_Y$ on Y are indicated. The central diagram portrays the situation when consideration is restricted to items in the population for which $a < X < b$. It is clear that for any particular X in this interval, the distribution and hence the mean $\bar{Y}_X$ of the corresponding Y's is the same as in the unrestricted case (top diagram). Consequently, a line of the form $\bar{Y}_X = b_0 + b_1X$ fitted to data involving either a random or selected set of values of X between $X = a$ and $X = b$, but with *no* selection or restrictions on the corresponding Y's, will furnish an unbiased estimate of the *true* regression line $\bar{Y}_X = \beta_0 + \beta_1X$ in the population at large. In contrast, if consideration is restricted to items for which $c < Y < d$, as indicated in the bottom diagram, then it is clear that the mean value, say $\bar{Y}'_X$, of the (restricted) Y's associated with any particular value of $X > m_X$ will be less than the corresponding mean value $\bar{Y}_X$ in the population as a whole. Likewise, if $X < m_X$, then the mean $\bar{Y}'_X$ of the corresponding (restricted) Y's will be greater than $\bar{Y}_X$ in the population as a whole. Consequently, a line of the form $\bar{Y}'_X = b_0 + b_1X$ fitted to data involving selection or restriction of Y's will *not* furnish an unbiased estimate of the true regression line $\bar{Y}_X = \beta_0 + \beta_1X$ in the population as a whole, and the distortion may be serious. In other words, introducing a restriction with regard to X does not bias inferences with regard to Y, when Y is considered as the dependent variable, but restricting Y will distort the dependence of $\bar{Y}_X$ on X so that the relationship observed will not be representative of the true underlying relationship in the population as a whole. Obviously, there is an equivalent statement in which the roles of X and Y are reversed. For further discussion and illustration of this point, and of the corresponding distortion of the sample correlation coefficient r as a measure of the true coefficient of correlation ρ in the populations, when *either* X *or* Y is restricted, see Eisenhart[1] and Ezekiel.[2]

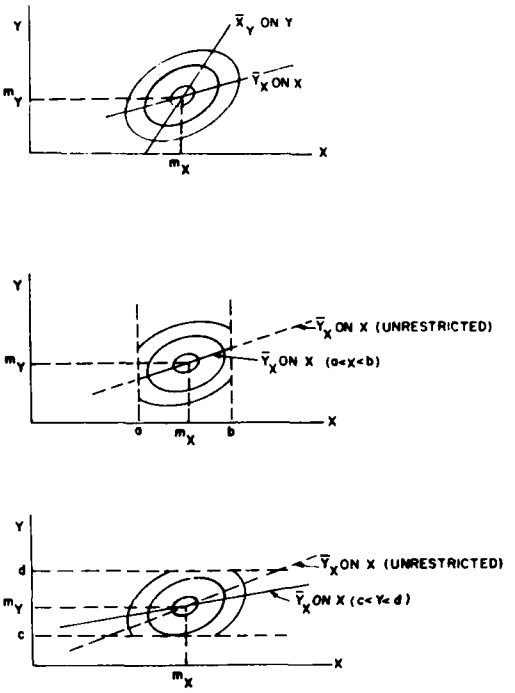


Figure 5-6. Diagram showing effect of restrictions of X or Y on the regression of Y on X.

As an engineering example of SII, consider a study of watches to investigate whether there was a relationship between the cost of a stop watch and its temperature coefficient. It was suggested that a correlation coefficient be computed. This was not possible because the watches had not been selected at random from the total watch production, but a deliberate effort had been made to obtain a fixed number of low-priced, medium-priced, and high-priced stop watches.

In any given case, consider carefully whether one is measuring samples as they come (and thereby accepting the values of both properties that come with the sample) which is an SI Relationship, or whether one selects samples which

TABLE 5-1. SUMMARY OF FOUR CASES OF LINEAR RELATIONSHIPS

	Functional (F)		Statistical (S)	
	FI	FII	SI	SII
Distinctive Features and Example	x and y are linearly related by a mathematical formula, $y = \beta_0 + \beta_1 x$, or $x = \beta_0' + \beta_1' y$, which is not observed exactly because of disturbances or errors in one or both variables. Example: Determination of elastic constant of a spring which obeys Hooke's law. x = accurately-known weight applied, Y = measured value of corresponding elongation y.		X = Height Y = Weight Both measured on a random sample of individuals. X is *not* selected but "comes with" sample unit.	X = Height (preselected values) Y = Weight of individuals of preselected height X is measured beforehand; only *selected* values of X are used at which to measure Y.
Errors of Measurement	Measurement error affects Y only.	X and Y both subject to error.	Ordinarily negligible compared to variation among individuals.	Same as in SI.
Form of Line Fitted	$Y = b_0 + b_1 x$	See Paragraph 5-4.3.	$\bar{Y}_X = b_0 + b_1 X$ $\bar{X}_Y = b_0' + b_1' Y$	$\bar{Y}_X = b_0 + b_1 X$ only.
Procedure for Fitting	See Paragraphs 5-4.1, 5-4.2, and basic worksheet.	Procedure depends on what assumptions can be made. See Paragraph 5-4.3.	See Paragraph 5-5.1 and basic worksheet.	See Paragraph 5-5.2 and basic worksheet.
Correlation Coefficient	Not applicable	Not applicable	Sample estimate is $r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$ See Paragraph 5-5.1.5.	Correlation may exist in the population, but r computed from *such* an experiment would provide a distorted estimate of the correlation.

are known to have a limited range of values of X (which is an SII Relationship).

Table 5-1 gives a brief summary characterization of SI and SII Relationships. Detailed problems and procedures with numerical examples are given for SI relationships in Paragraph 5-5.1 and for SII relationships in Paragraph 5-5.2.

BASIC WORKSHEET FOR ALL TYPES OF LINEAR RELATIONSHIPS

X denotes _____ Y denotes _____

ΣX = _____ ΣY = _____

$\bar{X}$ = _____ $\bar{Y}$ = _____

Number of points: n = _____

Step (1) ΣXY = _____

(2) $(\Sigma X)(\Sigma Y)/n$ = _____

(3) S_{xy} = Step (1) − Step (2)

(4) ΣX^2 = _____ (7) ΣY^2 = _____

(5) $(\Sigma X)^2/n$ = _____ (8) $(\Sigma Y)^2/n$ = _____

(6) S_{xx} = Step (4) − Step (5) (9) S_{yy} = Step (7) − Step (8)

(10) $b_1 = \dfrac{S_{xy}}{S_{xx}}$ = Step (3) ÷ Step (6) (14) $\dfrac{(S_{xy})^2}{S_{xx}}$ = _____

(11) $\bar{Y}$ = _____ (15) $(n-2)\,s_Y^2$ = Step (9) − Step (14)

(12) $b_1\bar{X}$ = _____ (16) s_Y^2 = Step (15) ÷ $(n-2)$

(13) $b_0 = \bar{Y} - b_1\bar{X}$ = Step (11) − Step (12) s_Y = _____

Equation of the line:

$Y = b_0 + b_1X$

_____

s_{b_1} = _____

s_{b_0} = _____

Estimated variance of the slope:

$s_{b_1}^2 = \dfrac{s_Y^2}{S_{xx}}$ = Step (16) ÷ Step (6)

Estimated variance of intercept:

$s_{b_0}^2 = s_Y^2\left\{\dfrac{1}{n} + \dfrac{\bar{X}^2}{S_{xx}}\right\}$ = _____

Note: The following are algebraically identical:

$S_{xx} = \Sigma(X - \bar{X})^2$; $S_{yy} = \Sigma(Y - \bar{Y})^2$; $S_{xy} = \Sigma(X - \bar{X})(Y - \bar{Y})$.

Ordinarily, in hand computation, it is preferable to compute as shown in the steps above. Carry all decimal places obtainable—i.e., if data are recorded to two decimal places, carry four places in Steps (1) through (9) in order to avoid losing significant figures in subtraction.

5-4 PROBLEMS AND PROCEDURES FOR FUNCTIONAL RELATIONSHIPS

5-4.1 FI RELATIONSHIPS (General Case)

There is an underlying mathematical (functional) relationship between the two variables, of the form $y = \beta_0 + \beta_1 x$. The variable x can be measured relatively accurately. Measurements Y of the value of y corresponding to a given x follow a normal distribution with mean $\beta_0 + \beta_1 x$ and variance $\sigma^2_{Y \cdot x}$ which is independent of the value of x. Furthermore, we shall assume that the deviations or *errors* of a series of observed Y's, corresponding to the same or different x's, all are mutually independent. See Paragraph 5-3.1 and Table 5-1.

The general case is discussed here, and the special case where it is known that $\beta_0 = 0$ (i.e., a line known to pass through the origin) is discussed in Paragraph 5-4.2. The procedure discussed here also will be valid if in fact $\beta_0 = 0$ even though this fact is not known beforehand. However, when it is known that $\beta_0 = 0$, the procedures of Paragraph 5-4.2 should be followed because they are simpler and somewhat more efficient.

It will be noted that SII, Paragraph 5-5.2, is handled computationally in exactly the same manner as FI, but both the underlying assumptions and the interpretation of the end results are different.

Data Sample 5-4.1—Young's Modulus vs. Temperature for Sapphire Rods

Observed values (Y) of Young's modulus (y) for sapphire rods measured at different temperatures (x) are given in the following table. There is assumed to be a linear functional relationship between the two variables x and y. (For the purpose of computation, the observed Y values were coded by subtracting 4000 from each. To express the line in terms of the original units, add 4000 to the computed intercept; the slope will not be affected.) The observed data are plotted in Figure 5-7.

x = Temperature °C	Y = Young's Modulus	Coded Y = Young's Modulus minus 4000
30	4642	642
100	4612	612
200	4565	565
300	4513	513
400	4476	476
500	4433	433
600	4389	389
700	4347	347
800	4303	303
900	4251	251
1000	4201	201
1100	4140	140
1200	4100	100
1300	4073	73
1400	4024	24
1500	3999	−1

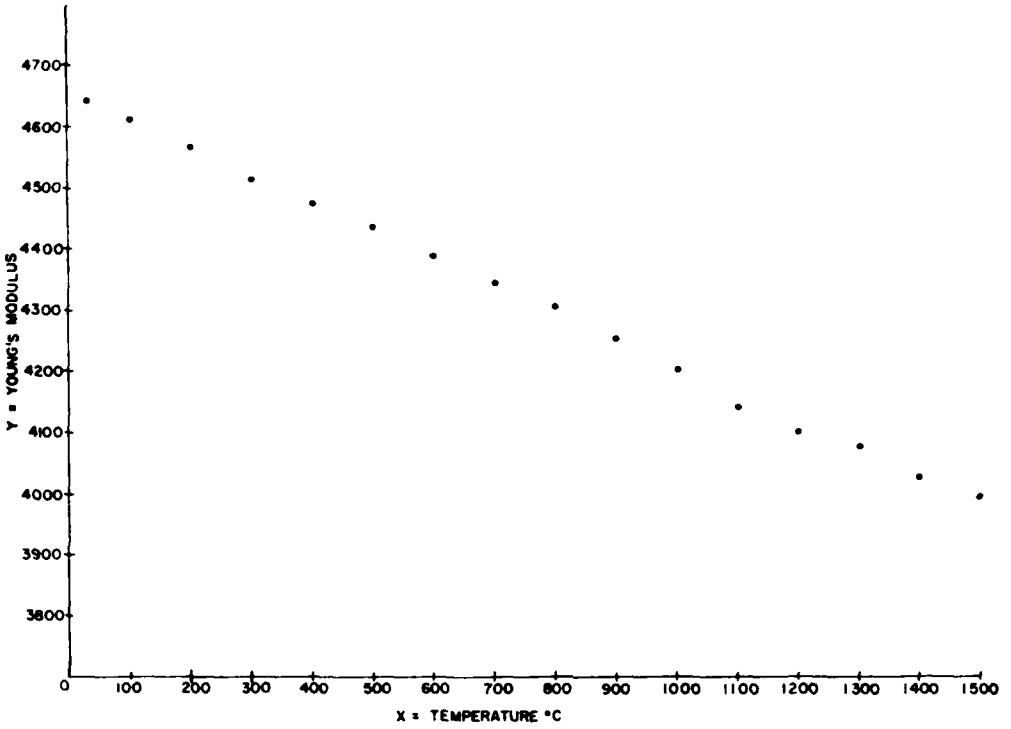


Figure 5-7. Young's modulus of sapphire rods as a function of temperature—an FI relationship.

5-4.1.1 What is the Best Line to be Used for Estimating y From Given Values of x?

CAUTION: Extrapolation, i.e., use of the line for prediction outside the range of data from which the line was computed, may lead to highly erroneous conclusions.

Procedure

Using Worksheet (See Worksheet 5-4.1), compute the line $Y = b_0 + b_1x$. This is an estimate of the true equation $y = \beta_0 + \beta_1x$. The method of fitting a line given here is a particular application of the general method of least squares. From Data Sample 5-4.1, the equation of the fitted line (in original units) is:

$$Y = 4654.9846 - 0.44985482\,x.$$

The equation in original units is obtained by adding 4000 to the computed intercept b_0. Since the Y's were coded by subtracting a constant, the computed slope b_1 was not affected. In Figure 5-8, the line is drawn and confidence limits for the line (computed as described in Paragraph 5-4.1.2.1) also are shown.

WORKSHEET 5-4.1
EXAMPLE OF FI RELATIONSHIP
YOUNG'S MODULUS AS FUNCTION OF TEMPERATURE

X denotes	Temperature, °C	Y denotes	Young's Modulus − 4000
ΣX =	12030	ΣY =	5068
$\bar{X}$ =	751.875	$\bar{Y}$ =	316.75

Number of points: n = 16

(1) ΣXY = 2,300,860

(2) $(\Sigma X)(\Sigma Y)/n$ = 3,810,502.5

(3) S_{xy} = −1,509,642.5

(4) ΣX^2 =	12,400,900	(7) ΣY^2 =	2,285,614
(5) $(\Sigma X)^2/n$ =	9,045,056.25	(8) $(\Sigma Y)^2/n$ =	1,605,289.
(6) S_{xx} =	3,355,843.75	(9) S_{yy} =	680,325.

(10) $b_1 = \frac{S_{xy}}{S_{xx}}$ =	−.449,854,82	(14) $\frac{(S_{xy})^2}{S_{xx}}$ =	679,119.9614
(11) $\bar{Y}$ =	316.75	(15) $(n-2)\,s_Y^2$ =	1,205.0386
(12) $b_1\bar{X}$ =	−338.2346	(16) s_Y^2 =	86.074 1857
(13) $b_0 = \bar{Y} - b_1\bar{X}$ =	654.9846	s_Y =	9.277617

b_0 (in original units) = 4654.9846

Equation of the line:
(in original units)

$Y = b_0 + b_1X$

Y = 4654.9846 − .449,854,82 x

s_{b_1} = .005 064

s_{b_0} = 4.458 638

Estimated variance of the slope:

$s_{b_1}^2 = \frac{s_Y^2}{S_{xx}}$ = .000 025 649 045

Estimated variance of intercept:

$s_{b_0}^2 = s_Y^2\left\{\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right\}$ = 19.879 452

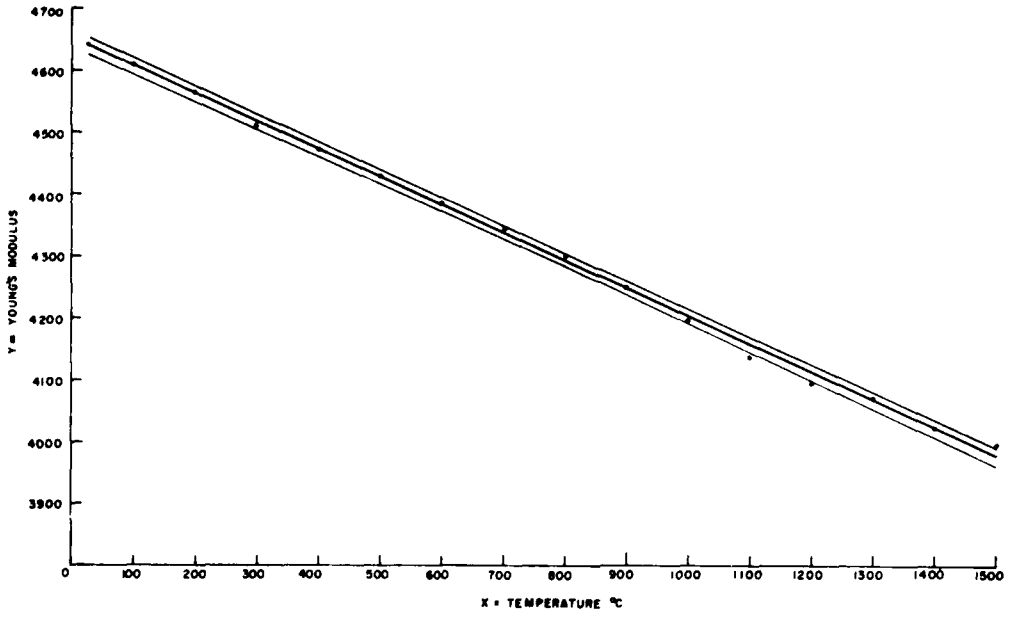


Figure 5-8. Young's modulus of sapphire rods as a function of temperature—showing computed regression line and confidence interval for the line.

Using the Regression Equation for Prediction. The fitted regression equation may be used for two kinds of predictions:

(a) To estimate the *true* value of y associated with a particular value of x, e.g., given $x = x'$ to estimate the value of $y' = \beta_0 + \beta_1 x'$; or,

(b) To predict a single new observed value Y corresponding to a particular value of x, e.g., given $x = x'$ to predict the value of a single measurement of y'.

Which prediction should be made? In some cases, it is sufficient to say that the *true* value of y (for given x) lies in a certain interval, and in other cases we may need to know how large (or how small) an individual observed Y value is likely to be associated with a particular value of x. The question of what to predict is similar to the question of what to specify (e.g., whether to specify average tensile strength or to specify minimum tensile strength) and can be answered only with respect to a particular situation. The difference is that here we are concerned with relationships between two variables and therefore must always talk about the value of y, or Y, for fixed x.

The predicted y' or Y' value is obtained by substituting the chosen value (x') of x in the fitted equation. For a particular value of x, either type of prediction ((a) or (b)) gives the same numerical answer for y' or Y'. The uncertainty associated with the prediction, however, does depend on whether we are estimating the *true* value of y', or predicting the value Y' of an individual measurement of y'. If the experiment could be repeated many times, each time obtaining n pairs of (x, Y) values, consider the range of Y values which would be obtained for a given x. Surely the individual Y values in all the sets will spread over a larger range than will the collection consisting of the average Y's (one from each set).

To estimate the *true* value of y associated with the value x', use the equation

$$y'_c = b_0 + b_1 x'.$$

The variance of y'_c as an estimate of the *true* value $y' = \beta_0 + \beta_1 x'$ is

$$\text{Var } y'_c = s^2_{Y \cdot x}\left[\frac{1}{n} + \frac{(X' - \bar{X})^2}{S_{xx}}\right]$$

This variance is the variance of estimate of a point on the fitted line.

For example, using the equation relating Young's modulus to temperature, we predict a value for y at $x = 1200$:

$$y'_c = 4654.9846 - .44985482\,(1200)$$
$$y'_c = 4115.16$$
$$\text{Var } y'_c = 86.074\left[.0625 + \frac{(1200 - 751.875)^2}{3{,}355{,}843.75}\right]$$
$$= 86.074\,(.0625 + .0598)$$
$$= 86.074\,(.1223)$$
$$\text{Var } y'_c = 10.53$$

To predict a single observed value of Y corresponding to a given value (x') of x, use the same equation

$$Y'_c = b_0 + b_1 x'.$$

The variance of Y'_c as an estimate of a single new (additional, future) measurement of y' is

$$\text{Var } Y'_c = s^2_{Y \cdot x}\left[1 + \frac{1}{n} + \frac{(X' - \bar{X})^2}{S_{xx}}\right]$$

The equation for our example is

$$Y = 4654.9846 - .44985482\,x.$$

To predict the value of a single determination of Young's modulus at $x = 750$, substitute in this equation and obtain:

$$Y'_c = 4654.9846 - .44985482\,(750)$$
$$= 4317.59$$
$$\text{Var } Y'_c = s^2_Y\left[1 + \frac{1}{n} + \frac{(X' - \bar{X})^2}{S_{xx}}\right]$$
$$= 86.074\left[1 + .0625 + \frac{(750 - 751.875)^2}{3{,}355{,}843.75}\right]$$
$$= 86.074\,(1.0625)$$
$$= 91.45\ .$$

5-4.1.2 What are the Confidence Interval Estimates for: the Line as a Whole; a Point on the Line; a Future Value of Y Corresponding to a Given Value of x?

Once we have fitted the line, we want to make predictions from it, and we want to know how good our predictions are. Often, these predictions will be given in the form of an interval together with a confidence coefficient associated with the interval—i.e., confidence interval estimates. Several kinds of confidence interval estimates may be made:

(a) A confidence band for the line as a whole.

(b) A confidence interval for a point on the line—i.e., a confidence interval for y' (the *true* value of y and the *mean* value of Y) corresponding to a single value of $x = x'$.

If the fitted line is, say, a calibration line which will be used over and over again, we will want to make the interval estimate described in (a). In other cases, the line as such may not be so important. The line may have been fitted only to investigate or check the structure of the relationship, and the interest of the experimenter may be centered at one or two values of the variables.

Another kind of interval estimate sometimes is required:

(c) A single observed value (Y') of Y corresponding to a new value of $x = x'$.

These three kinds of interval statements have somewhat different interpretations. The confidence interval for (b) is interpreted as follows:

Suppose that we repeated our experiment a large number of times. *Each time*, we obtain n pairs of values (x_i, Y_i), fit the line, and compute a confidence interval estimate for $y' = \beta_0 + \beta_1 x'$, the value of y corresponding to the particular value $x = x'$. Such interval estimates of y' are expected to be correct (i.e., include the *true* value of y') a proportion $(1 - \alpha)$ of the time. If we were to make an interval estimate of y'' corresponding to another value of $x = x''$, these interval estimates also would be expected to include y'' the same proportion $(1 - \alpha)$ of the time. However, taken together, these intervals do not constitute a joint confidence statement about y' and y'' which would be expected to be correct exactly a proportion $(1 - \alpha)$ of the

time; nor is the effective level of confidence $(1 - \alpha)^2$, because the two statements are not independent but are correlated in a manner intimately dependent on the values x' and x'' for which the predictions are to be made.

The confidence band for the whole line (a) implies the same sort of repetition of the experiment except that our confidence statements are not now limited to one x at a time, but we can talk about any number of x values simultaneously—about the whole line. Our confidence statement applies to the line as a whole, and therefore the confidence intervals for y corresponding to all the chosen x values will simultaneously be correct a proportion $(1 - \alpha)$ of the time. It will be noted that the intervals in (a) are larger than the intervals in (b) by the ratio $\sqrt{2F}/t$. This wider interval is the "price" we pay for making joint statements about y for any number of or for all of the x values, rather than the y for a single x.

Another *caution* is in order. We cannot use the same computed line in (b) and (c) to make a large number of predictions, and claim that 100 $(1 - \alpha)$ % of the predictions will be correct. The *estimated* line may be very close to the *true line*, in which case nearly all of the interval predictions may be correct; or the line may be considerably different from the *true line*, in which case very few may be correct. In practice, provided our situation is *in control*, we should always revise our estimate of the line to include additional information in the way of new points.

5-4.1.2.1 What is the $(1 - \alpha)$ Confidence Band for the Line as a Whole?

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Obtain s_Y from Worksheet.	(2) $s_Y = 9.277617$ from Worksheet 5-4.1
(3) Look up $F_{1-\alpha}$ for $(2, n - 2)$ degrees of freedom in Table A-5.	(3) $F_{.95}(2, 14) = 3.74$
(4) Choose a number of values of X (within the range of the data) at which to compute points for drawing the confidence band.	(4) Let: $X = 30$ $X = 400$ $X = 800$ $X = 1200$ $X = 1500$, for example.
(5) At each selected value of X, compute: $Y_c = \bar{Y} + b_1 (X - \bar{X})$ and $W_1 = \sqrt{2F}\, s_Y \left[\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{xx}}\right]^{\frac{1}{2}}$	(5) See Table 5-2 for a convenient computational arrangement and the example calculations.
(6) A $(1 - \alpha)$ confidence band for the whole line is determined by $Y_c \pm W_1$.	(6) See Table 5-2.

Procedure

(7) To draw the line and its confidence band, plot Y_c at two of the extreme selected values of X. Connect the two points by a straight line. At each selected value of X, also plot $Y_c + W_1$ and $Y_c - W_1$. Connect the upper series of points, and the lower series of points, by smooth curves.

If more points are needed for drawing the curves for the band, note that, because of symmetry, the calculation of W_1 at n values of X actually gives W_1 at $2n$ values of X.

Example

(7) See Figure 5-8.

For example: W_1 (but not Y_c) has the same value at $X = 400$ (i.e., $\bar{X} - 351.875$) as at $X = 1103.75$ (i.e., $\bar{X} + 351.875$).

TABLE 5-2. COMPUTATIONAL ARRANGEMENT FOR PROCEDURE 5-4.1.2.1

X	$(X - \bar{X})$	Y_c	$\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{xx}}$	$s^2_{Y_c}$	s_{Y_c}	W_1	$Y_c + W_1$	$Y_c - W_1$
30	−721.875	4641.49	.21778	18.7452	4.3296	11.84	4653.33	4629.65
400	−351.875	4475.04	.09940	8.5558	2.9250	8.00	4483.04	4467.04
800	48.125	4295.10	.06319	5.4390	2.3322	6.38	4301.48	4288.72
1200	448.125	4115.16	.12234	10.5303	3.2450	8.88	4124.04	4106.28
1500	748.125	3980.20	.22928	19.7351	4.4424	12.15	3992.35	3968.05

$\bar{X} = 751.875$

coded $\bar{Y} = 316.75$

$\bar{Y}$ (original units) $= 4316.75$

$s^2_Y = 86.0741857$

$\frac{1}{n} = .0625$

$b_1 = -.44985482$

$S_{xx} = 3{,}355{,}843.75$

$\sqrt{2F} = 2.735$

$Y_c = \bar{Y} + b_1 (X - \bar{X})$

$s^2_{Y_c} = s^2_Y \left[\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{xx}}\right]$

$W_1 = 2.735\, s_{Y_c}$

5-4.1.2.2 Give a $(1 - \alpha)$ Confidence Interval Estimate fo. a Single Point on the Line (i.e., the Mean Value of Y Corresponding to a Chosen Value of $x = x'$)

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Obtain s_Y from Worksheet.	(2) $s_Y = 9.277617$ from Worksheet 5-4.1
(3) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(3) $t_{.975}\,(14) = 2.145$
(4) Choose X', the value of X at which we want to make an interval estimate of the mean value of Y.	(4) Let $X' = 1200$
(5) Compute: $W_2 = t_{1-\alpha/2}\, s_Y \left[\frac{1}{n} + \frac{(X' - \bar{X})^2}{S_{xx}}\right]^{\frac{1}{2}}$ and $Y_c = \bar{Y} + b_1 (X' - \bar{X})$	(5) $W_2 = 2.145\ (3.2451)$ $= 6.96$ $Y_c = 4115.16$
(6) A $(1 - \alpha)$ confidence interval estimate for the mean value of Y corresponding to $X = X'$ is given by $Y_c \pm W_2$.	(6) A 95% confidence interval estimate for the mean value of Y corresponding to $X = 1200$ is 4115.16 ± 6.96 $= 4108.20$ to 4122.12.

Note: An interval estimate of the intercept of the line (β_0) is obtained by setting $X' = 0$ in the above procedure.

5-4.1.2.3 Give a $(1 - \alpha)$ Interval Estimate for a Single (Future) Value (Y') of Y Corresponding to a Chosen Value (x') of x.

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Obtain s_Y from Worksheet.	(2) $s_Y = 9.277617$ from Worksheet 5-4.1
(3) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(3) $t_{.975}\,(14) = 2.145$
(4) Choose X', the value of X at which we want to make an interval estimate of a single value of Y.	(4) Let $X' = 1200$
(5) Compute: $W_3 = t_{1-\alpha/2}\, s_Y \left[1 + \frac{1}{n} + \frac{(X' - \bar{X})^2}{S_{xx}}\right]^{\frac{1}{2}}$ and $Y_c = \bar{Y} + b_1 (X' - \bar{X})$	(5) $W_3 = 2.145\ (9.8288)$ $= 21.08$ $Y_c = 4115.16$
(6) A $(1 - \alpha)$ confidence interval estimate for Y' (the single value of Y corresponding to X') is $Y_c \pm W_3$.	(6) A 95% confidence interval estimate for a single value of Y corresponding to $X' = 1200$ is 4115.16 ± 21.08 $= 4094.08$ to 4136.24 .

5-4.1.3 What is the Confidence Interval Estimate for β_1, the Slope of the True Line $y = \beta_0 + \beta_1 x$?

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(2) $t_{.975}\,(14) = 2.145$
(3) Obtain s_{b_1} from Worksheet.	(3) $s_{b_1} = .005064$ from Worksheet 5.4.1
(4) Compute $W_4 = t_{1-\alpha/2}\, s_{b_1}$	(4) $W_4 = 2.145\ (.005064)$ $= .010862$
(5) A $(1 - \alpha)$ confidence interval estimate for β_1 is $b_1 \pm W_4$.	(5) $b_1 = -.449855$ $W_4 = .010862$ A 95% confidence interval for β_1 is the interval $-.449855 \pm .010862$, i.e., the interval from $-.460717$ to $-.438993$.

5-4.1.4 If We Observe n' New Values of Y (with Average $\bar{Y}'$), How Can We Use the Fitted Regression Line to Obtain an Interval Estimate of the Value of x that Produced These Values of Y?

Example: Suppose that we obtain 10 new measurements of Young's modulus (with average, $\bar{Y}' = 4500$) and we wish to use the regression line to make an interval estimate of the temperature (x) at which the measurements were made.

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(2) $t_{.975}\,(14) = 2.145$
(3) Obtain b_1 and $s_{b_1}^2$ from Worksheet.	(3) From Worksheet 5-4.1, $b_1 = -.449855$ $s_{b_1}^2 = .0000256490$
(4) Compute $C = b_1^2 - (t_{1-\alpha/2})^2 s_{b_1}^2$	(4) $C = .202370 - .000118$ $= .202252$
(5) A $(1 - \alpha)$ confidence interval estimate for the X corresponding to $\bar{Y}'$ is computed from $X' = \bar{X} + \frac{b_1 (\bar{Y}' - \bar{Y})}{C} \pm \frac{t_{1-\alpha/2}\, S_Y}{C} \sqrt{\frac{(\bar{Y}' - \bar{Y})^2}{S_{xx}} + \left(\frac{1}{n} + \frac{1}{n'}\right) C}$	(5) A 95% confidence interval would be computed as follows: $X' = 751.875 - \frac{.449855\,(4500 - 4316.75)}{.202252} \pm \frac{2.145\,(9.277617)}{.202252} \times \sqrt{\frac{(183.25)^2}{3{,}355{,}843.75} + (.1625)\,(.202252)}$ $= 751.875 - 407.590 \pm 98.39452 \sqrt{.0100066 + .0328660}$ $= 344.285 \pm 98.39452 \sqrt{.0428726}$ $= 344.285 \pm 98.39452\,(.20706)$ $= 344.285 \pm 20.374$

The interval from $X = 323.911$ to $X = 364.659$ is a 95% confidence interval for the value of temperature which produced the 10 measurements whose mean Young's modulus was 4500.

5-4.1.5 Using the Fitted Regression Line, How Can We Choose a Value (x') of x Which We May Expect with Confidence ($1 - \alpha$) Will Produce a Value of Y Not Less Than Some Specified Value Q?

Example: What value (x') of temperature (x) can be expected to produce a value of Young's modulus not less than 4300?

Procedure

(1) Choose the desired confidence level, $1 - \alpha$; and choose Q

(2) Look up $t_{1-\alpha}$ for $n - 2$ degrees of freedom in Table A-4.

(3) Obtain b_1 and $s_{b_1}^2$ from Worksheet.

(4) Compute

$$C = b_1^2 - (t_{1-\alpha})^2 s_{b_1}^2$$

(5) Compute

$$X' = \bar{X} + b_1\left(\frac{Q - \bar{Y}}{C}\right) \pm \frac{t_{1-\alpha} s_Y}{C}\sqrt{\frac{(Q - \bar{Y})^2}{S_{xx}} + \left(\frac{n+1}{n}\right)C}$$

where the sign before the last term is + if b_1 is positive or − if b_1 is negative. We have confidence $(1 - \alpha)$ that a value of $X = X'$ will correspond to (produce) a value of Y not less than Q. (See discussion of "confidence" in straight-line prediction in Paragraph 5-4.1.2).

Example

(1) Let: $1 - \alpha = .95$
$\alpha = .05$
$Q = 4300$

(2) $t_{.95}\,(14) = 1.761$

(3) From Worksheet 5-4.1,
$b_1 = -.449855$
$s_{b_1}^2 = .0000256490$

(4)

$$C = .202370 - .000080 = .202290$$

(5) The value of X' is computed as follows:

$$X' = 751.875 + \frac{-.449855\,(4300 - 4316.75)}{.202290} - \frac{1.761\,(9.277617)}{.202290} \times \sqrt{\frac{(4300 - 4316.75)^2}{3{,}355{,}843.75} + \left(\frac{17}{16}\right)C}$$

$$= 751.875 + 37.249 - 80.764662\sqrt{.000084 + .214933}$$

$$= 751.875 + 37.249 - 80.764662\sqrt{.215017}$$

$$= 751.875 + 37.249 - 37.450$$

$$= 751.674\ .$$

5-4.1.6 Is the Assumption of Linear Regression Justified?

This involves a test of the assumption that the mean Y values ($\bar{Y}_x$) for given x values do lie on a straight line (we assume that for any given value of x, the corresponding individual Y values are normally distributed with variance σ_Y^2, which is independent of the value of x). A simple test is available provided that we have more than one observation on Y at one or more values of x. Assume that there are n pairs of values (x_i, Y_i), and that among these pairs there occur only k values of x (where k is less than n).

For example, see the data recorded in Table 5-3 which shows measurements of Young's modulus (coded) of sapphire rods as a function of temperature.

Each x is recorded in Column 1, and the corresponding Y values (varying in number from 1 to 3 in the example) are recorded opposite the appropriate x. The remaining columns in the table are convenient for the required computations.

TABLE 5-3. COMPUTATIONAL ARRANGEMENT FOR TEST OF LINEARITY

X = Temperature	Y = Young's Modulus Minus 3000			ΣY	$(\Sigma Y)^2$	ΣY^2	n_i	$n_i X_i$	$n_i X_i^2$	ΣXY	$\frac{(\Sigma Y)^2}{n_i}$
500	328			328	107584	107584	1	500	250000	164000	107584
550	296			296	87616	87616	1	550	302500	162800	87616
600	266			266	70756	70756	1	600	360000	159600	70756
603	260	244		504	254016	127136	2	1206	727218	303912	127008
650	240	232	213	685	469225	156793	3	1950	1267500	445250	156408.3
700	204	203	184	591	349281	116681	3	2100	1470000	413700	116427
750	174	175	154	503	253009	84617	3	2250	1687500	377250	84336.3
800	152	146	124	422	178084	59796	3	2400	1920000	337600	59361.3
850	117	94		211	44521	22525	2	1700	1445000	179350	22260.5
900	97	61		158	24964	13130	2	1800	1620000	142200	12482
950	38			38	1444	1444	1	950	902500	36100	1444
1000	30	5		35	1225	925	2	2000	2000000	35000	612.5
		TOTAL		4037 $= T_1$		849003 $= T_2$	24 $= n$	18006 $= T_3$	13952218 $= T_4$	2756762 $= T_5$	846296 $= T_6$

Procedure	Example
(1) Choose α, the significance level of the test.	(1) Let: $\alpha = .05$, $1 - \alpha = .95$
(2) Compute: $\bar{Y} = \frac{T_1}{n}$ $\bar{X} = \frac{T_3}{n}$, the weighted average of X.	(2) $\bar{Y} = \frac{4037}{24} = 168.21$ $\bar{X} = \frac{18006}{24} = 750.25$
(3) Compute $S_1 = T_6 - \frac{(T_1)^2}{n}$	(3) $\frac{(T_1)^2}{n} = 679057.04$ $S_1 = 846296 - 679057.04 = 167238.96$
(4) Compute $b = \frac{T_5 - \frac{T_3 T_1}{n}}{T_4 - \frac{(T_3)^2}{n}}$	(4) $b = \frac{2756762 - 3028759.25}{13952218 - 13509001.5} = \frac{-271997.25}{443216.5} = -0.6136894$
(5) Compute $S_2 = b\left(T_5 - \frac{T_3 T_1}{n}\right)$	(5) $S_2 = -0.6136894\ (-271997.25) = 166921.83$
(6) Compute $S_3 = T_2 - \frac{(T_1)^2}{n}$	(6) $S_3 = 849003 - 679057.04 = 169945.96$
(7) Look up $F_{1-\alpha}$ for $(k - 2, n - k)$ degrees of freedom in Table A-5.	(7) $n = 24$, $k = 12$ $F_{.95}$ for (10, 12) degrees of freedom = 2.75
(8) Compute $F = \left(\frac{S_1 - S_2}{S_3 - S_1}\right)\left(\frac{n - k}{k - 2}\right)$	(8) $F = \left(\frac{317.13}{2707}\right)\left(\frac{24 - 12}{10}\right) = (.11715)\ (1.2) = 0.14$
(9) If $F > F_{1-\alpha}$, decide that the "array means" $\bar{Y}_x$ do not lie on a straight line. If $F < F_{1-\alpha}$, the hypothesis of linearity is not disproved.	(9) Since F is less than $F_{1-\alpha}$, the hypothesis of linearity is not disproved.

5-4.2 FI RELATIONSHIPS WHEN THE INTERCEPT IS KNOWN TO BE EQUAL TO ZERO (LINES THROUGH THE ORIGIN)

In Paragraph 5-4.1, we assumed:

(a) that there is an underlying linear functional relationship between x and y of the form $y = \beta_0 + \beta_1 x$, with intercept β_0 and slope β_1 both different from zero;

(b) that our data consist of observed values $Y_1, Y_2, \ldots, Y_n$ of y, corresponding to accurately-known values $x_1, x_2, \ldots, x_n$ of x; and,

(c) that the Y's can be regarded as being independently and normally distributed with means equal to their respective *true* values (i.e, mean of $Y_i = \beta_0 + \beta_1 x_i$, $i = 1, 2, \ldots, n$) and constant variance $\sigma^2_{Y\cdot x} = \sigma^2$ for all x.

Furthermore, we gave: a procedure (Paragraph 5-4.1.2.2 with $X' = 0$) for determining confidence limits for β_0, and hence for testing the hypothesis that $\beta_0 = 0$, in the absence of prior knowledge of the value of β_1; and a procedure that is independent of the value of β_0 (Paragraph 5-4.1.3) for determining confidence limits for β_1, and hence for testing the hypothesis that $\beta_1 = 0$.

We now consider the analysis of data corresponding to an FI structural relationship when it is known that $y = 0$ when $x = 0$, so that the line must pass through the origin, i.e., *when it is known that* $\beta_0 = 0$. To begin with, we assume as in (b) and (c) above, that our data consist of observed values $Y_1, Y_2, \ldots, Y_n$, of a *dependent* variable y corresponding to accurately-known values $x_1, x_2, \ldots, x_n$ of the *independent* variable x and that these Y's can be regarded as being *independently* and normally distributed with means $\beta_1 x_1, \beta_1 x_2, \ldots, \beta_1 x_n$, respectively, and variances $\sigma^2_{Y\cdot x}$ that may depend on x. We consider explicitly the cases of constant variance ($\sigma^2_{Y\cdot x} = \sigma^2$), variance proportional to x ($\sigma^2_{Y\cdot x} = x\sigma^2$), and standard deviation proportional to x ($\sigma_{Y\cdot x} = x\sigma$). Finally, we consider briefly the case of *cumulative data* where $x_1 < x_2 < \ldots < x_n$ and the error in Y_i is of the form $e_1 + e_2 + \ldots + e_{i-1} + e_i$, that is, is the sum of the errors of all preceding Y's plus a "private error" e_i of its own. Following Mandel,[3] we assume that the errors (e_i) are independently and normally distributed with zero means and with variances proportional to the length of their generation intervals, i.e., $\sigma^2_{e_i} = (x_i - x_{i-1})\sigma^2$. Under these circumstances, the Y's will be normally distributed with means $\beta_1 x_1, \beta_1 x_2, \ldots, \beta_1 x_n$, respectively, as before; and with variances $\sigma^2_{Y_i} = x_i\sigma^2$, respectively; but will not be independent owing to the overlap among their respective errors.

5-4.2.1 Line Through Origin, Variance of Y's Independent of x. The slope of the best-fitting line of the form $Y = b_1 x$ is given by

$$b_1 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

and the estimated variance of b_1 is

$$s^2_{b_1} = \frac{s^2_Y}{\sum_{i=1}^{n} x_i^2}$$

where

$$s^2_Y = \frac{\sum_{i=1}^{n} (Y_i - b_1 x_i)^2}{n-1} = \frac{\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{1}^{n} x_i Y_i\right)^2}{\sum_{1}^{n} x_i^2}}{n-1} .$$

Consequently, we may effect a simplification of our Basic Worksheet—see Worksheet 5-4.2.1.

Using the values of b_1 and s_{b_1} so obtained, confidence limits for β_1, the slope of the true line through the origin, $y = \beta_1 x$, can be obtained by following the procedure of Paragraph 5-4.1.3 using $t_{1-\alpha/2}$ for $n - 1$ degrees of freedom. Confidence limits for the line as a whole then are obtained simply by plotting the lines $y = \beta_1^U x$ and $y = \beta_1^L x$, where β_1^U and β_1^L are the upper and lower confidence limits for β_1 obtained in the manner just described. The limiting lines, in this instance, also furnish confidence limits for the value y' of y corresponding to a particular point on the line, say for $x = x'$, so that an additional procedure is unnecessary. Confidence limits for a single future observed Y corresponding to $x = x'$ are given by

$$b_1 x' \pm t_{1-\alpha/2} \sqrt{s^2_Y + (x')^2 s^2_{b_1}} ,$$

where s^2_Y and s_{b_1} are from our modified worksheet and $t_{1-\alpha/2}$ corresponds to $n - 1$ degrees of freedom.

WORKSHEET 5-4.2.1

WORKSHEET FOR FI RELATIONSHIPS WHEN THE INTERCEPT IS KNOWN TO BE ZERO AND THE VARIANCES OF THE Y's IS INDEPENDENT OF x

X denotes _____ Y denotes _____

ΣX = _____ ΣY = _____

$\bar{X}$ = _____ $\bar{Y}$ = _____

Number of points: n = _____

Step (1) ΣXY = _____

(2) ΣX^2 = _____ (5) $\dfrac{(\Sigma XY)^2}{\Sigma X^2}$ = _____

(3) ΣY^2 = _____ (6) $(n-1)\,s_Y^2$ = Step (3) − Step (5)

(4) $b_1 = \dfrac{\Sigma XY}{\Sigma X^2}$ = Step (1) ÷ Step (2) (7) s_Y^2 = Step (6) ÷ $(n-1)$

s_Y = _____

Equation of the Line:

$Y = b_1 X$

_____

Estimated variance of the slope:

$s_{b_1}^2 = \dfrac{s_Y^2}{\Sigma X^2}$ = Step (7) ÷ Step (2)

s_{b_1} = _____

5-4.2.2 Line Through Origin, Variance Proportional to x ($\sigma_{Y \cdot x}^2 = x\sigma^2$). The slope of the best-fitting line of form $Y = b_1 x$ is given by

$$b_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{Y}}{\bar{x}},$$

the *ratio of the averages*, and the estimated variance of b_1 is

$$s_{b_1}^2 = \frac{s^2}{\sum_{i=1}^{n} x_i}$$

where

$$(n-1)\,s^2 = \sum_{i=1}^{n} \left(\frac{Y_i^2}{x_i}\right) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{\sum_{i=1}^{n} x_i}$$

Using the values of b_1 and s_{b_1} so obtained, confidence limits for β_1, the slope of the true line through the origin, $y = \beta_1 x$, can be obtained by following the procedure of Paragraph 5-4.1.3 using $t_{1-\alpha/2}$ for $n-1$ degrees of freedom. Confidence limits for the line as a whole then are obtained simply by plotting the lines $y = \beta_1^U x$ and $y = \beta_1^L x$ where β_1^U and β_1^L are the upper and lower confidence limits for β_1 obtained in the manner just described. The limiting lines, in this instance, also furnish confidence limits for the value y' corresponding to a particular point on the line, say for $x = x'$. Confidence limits for a single future observed Y corresponding to $x = x'$, are given by

$$b_1 x' \pm t_{1-\alpha/2} \sqrt{x' s^2 + (x')^2 s_{b_1}^2},$$

where s_{b_1} is computed as shown above and $t_{1-\alpha/2}$ corresponds to $n-1$ degrees of freedom.

5-4.2.3 Line Through Origin, Standard Deviation Proportional to x ($\sigma_{Y \cdot x} = x\sigma$). The slope of the best-fitting line of form $Y = b_1 x$ is given by

$$b_1 = \sum_{i=1}^{n} \left(\frac{Y_i}{x_i}\right)/n,$$

the average of the ratios $\left(\frac{Y_i}{x_i}\right)$,

and the estimated variance of b_1 is

$$s_{b_1}^2 = \frac{s^2}{n}$$

where

$$(n-1)\, s^2 = \sum_{i=1}^{n} \left(\frac{Y_i}{x_i}\right)^2 - \frac{\left[\sum_{i=1}^{n} \left(\frac{Y_i}{x_i}\right)\right]^2}{n}$$

that is,

$$s^2 = \frac{\sum_{i=1}^{n} R_i - \frac{(\Sigma R_i)^2}{n}}{n(n-1)}$$

for $$R_i = \frac{Y_i}{x_i}$$

Using the values of b_1 and s_{b_1} so obtained, confidence limits for β_1, the slope of the true line through the origin, $y = \beta_1 x$, can be obtained by following the procedure of Paragraph 5-4.1.3 using $t_{1-\alpha/2}$ for $n-1$ degrees of freedom. Confidence limits for the line as a whole are then obtained simply by plotting the lines $y = \beta_1^U x$ and $y = \beta_1^L x$ where β_1^U and β_1^L are the upper and lower confidence limits for β_1 obtained in the manner just described. The limiting lines, in this instance, also furnish confidence limits for the value y' of y corresponding to a particular point on the line, say for $x = x'$. Confidence limits for a single future observed Y corresponding to $x = x'$, are given by

$$b_1 x' \pm t_{1-\alpha/2}\, x' \sqrt{s^2 + s_{b_1}^2}\ ,$$

where s_{b_1} is computed as shown above and $t_{1-\alpha/2}$ corresponds to $n-1$ degrees of freedom.

5-4.2.4 Line Through Origin, Errors of Y's Cumulative (Cumulative Data). In many engineering tests and laboratory experiments the observed values $Y_1, Y_2, \ldots, Y_i, \ldots$, of a dependent variable y represent the cumulative magnitude of some effect at successive values $x_1 < x_2 < x_3 < \ldots$ of the independent variable x. Thus, $Y_1, Y_2, \ldots$, may denote: the total weight loss of a tire under road test, measured at successive mileages $x_1, x_2, \ldots$; or the weight gain of some material due to water absorption at successive times $x_1, x_2, \ldots$; or the total deflection of a beam (or total compression of a spring) under continually increasing load, measured at loads $x_1, x_2, \ldots$; and so forth. In such cases, even though the underlying functional relationship takes the form of a line through the origin, $y = \beta x$, none of the procedures that we have presented thus far will be applicable, because of the cumulative effect of errors of technique on the successive Y's; the deviation of Y_i from its true or expected value y_i will include the deviation $(Y_{i-1} - y_{i-1})$ of Y_{i-1} from its true or expected value, plus an individual "private deviation or error" e_i of its own. Hence, the total error of Y_i will be the sum $(e_1 + e_2 + \ldots + e_{i-1} + e_i)$ of the individual error contributions of $Y_1, Y_2, \ldots, Y_{i-1}$, and its own additional deviation.

If the test or experiment starts at $x_0 = 0$, and the x's form an uninterrupted sequence $0 < x_1 < x_2 < \ldots < x_n$, and if we may regard the individual error contributions $e_1, e_2, \ldots$, as independently and normally distributed with zero means and variances proportional to the lengths of the x-intervals over which they accrue, i.e., if $\sigma_{e_i}^2 = (x_i - x_{i-1})\,\sigma^2$, then the best estimate of the slope of the underlying linear functional relation $y = \beta_1 x$ is given by

$$b_1 = \frac{Y_n}{x_n}$$

and estimated variance of b_1

$$s_{b_1}^2 = \frac{1}{(n-1)\, x_n} \left\{ \sum_{i=1}^{n} \frac{(Y_i - Y_{i-1})^2}{x_i - x_{i-1}} - \frac{Y_n^2}{x_n} \right\}$$

in which $x_0 = 0$ and $Y_0 = 0$ by hypothesis.

Using the values of b_1 and s_{b_1} so obtained, confidence limits for β_1, the slope of the true line through the origin, $y = \beta_1 x$, can be obtained by following the procedure of Paragraph 5-4.1.3 using $t_{1-\alpha/2}$ for $n-1$ degrees of freedom. Confidence limits for the line as a whole then are obtained simply by plotting the lines $y = \beta_1^U x$ and $y = \beta_1^L x$, where β_1^U and β_1^L are the upper and lower confidence limits for β_1 obtained in the manner just described. These limit lines also

provide confidence limits for a particular point on the line, say the value y' corresponding to $x = x'$. For the fitting of lines of this sort to cumulative data under more general conditions, and for other related matters, see Mandel's article.[3]

5-4.3 FII RELATIONSHIPS

Distinguishing Features. There is an underlying mathematical (functional) relationship between the two variables, of the form

$$y = \beta_0 + \beta_1 x.$$

Both X and Y are subject to errors of measurement. Read Paragraph 5-3.1 and Table 5-1.

The full treatment of this case depends on the assumptions we are willing to make about error distributions. For complete discussion of the problem, see Acton.[4]

5-4.3.1 A Simple Method of Fitting the Line In the General Case. There is a quick and simple method of fitting a line of the form $Y = b_0 + b_1 X$ which is generally applicable when both X and Y are subject to errors of measurement. This method is described in Bartlett,[5] and is illustrated in this paragraph. Similar methods had been used previously by other authors.

(a) For the location of the fitted straight line, use as the pivot point the center of gravity of all n observed points (X_i, Y_i), that is, the point with the mean coordinates $(\bar{X}, \bar{Y})$. In consequence, the fitted line will be of the form $Y = b_0 + b_1 X$ with $b_0 = \bar{Y} - b_1 \bar{X}$, just as in the least-squares method in Paragraph 5-4.1.

(b) For the slope, divide the n plotted points into three non-overlapping groups when considered in the X direction. There should be an equal number of points, k, in each of the two extreme groups, with k as close to $\frac{n}{3}$ as possible.

Take, as the slope of the line,

$$b_1 = \frac{\bar{Y}_3 - \bar{Y}_1}{\bar{X}_3 - \bar{X}_1},$$

where

$\bar{Y}_3$ = average Y for 3rd group
$\bar{Y}_1$ = average Y for 1st group
$\bar{X}_3$ = average X for 3rd group
$\bar{X}_1$ = average X for 1st group.

Data Sample 5-4.3.1—Relation of Two Colorimetric Methods

The following data are coded results of two colorimetric methods for the determination of a chemical constituent. (The data have been coded for a special purpose which has nothing to do with this illustration). The interest here, of course, is in the relationship between results given by the two methods, and it is presumed that there is a functional relationship with both methods subject to errors of measurement.

Sample	Method I X	Method II Y
1	3720	5363
2	4328	6195
3	4655	6428
4	4818	6662
5	5545	7562
6	7278	9184
7	7880	10070
8	10085	12519
9	11707	13980

(a) The fitted line must pass through the point $(\bar{X}, \bar{Y})$, where

$$\bar{X} = 6668.4$$
$$\bar{Y} = 8662.6$$

(b) To determine the slope, divide the points into 3 groups. Since there are 9 points, exactly 3 equal groups are obtained.

$$\bar{Y}_3 = 12190$$
$$\bar{Y}_1 = 5995$$
$$\bar{X}_3 = 9891$$
$$\bar{X}_1 = 4234$$
$$b_1 = \frac{\bar{Y}_3 - \bar{Y}_1}{\bar{X}_3 - \bar{X}_1} = \frac{12190 - 5995}{9891 - 4234} = \frac{6195}{5657} = 1.0951$$
$$b_0 = \bar{Y} - b_1 \bar{X} = 8662.6 - \frac{6195}{5657}(6668.4) = 1360.0 .$$

The fitted line

$$Y = 1360.0 + 1.0951\ X$$

is shown in Figure 5-9.

Procedures are given in Bartlett[8] for determining 100 $(1 - \alpha)$ % confidence limits for the true slope β_1; and for determining a 100 $(1 - \alpha)$ % confidence ellipse for β_0 and β_1 jointly, from which 100 $(1 - \alpha)$ % confidence limits for the line as a whole can be derived. For strict validity, they require that the measurement errors affecting the observed X_i be sufficiently small in comparison with the spacing of their true values x_i that the allocation of the observational points (X_i, Y_i) to the three groups is unaffected. These procedures are formally

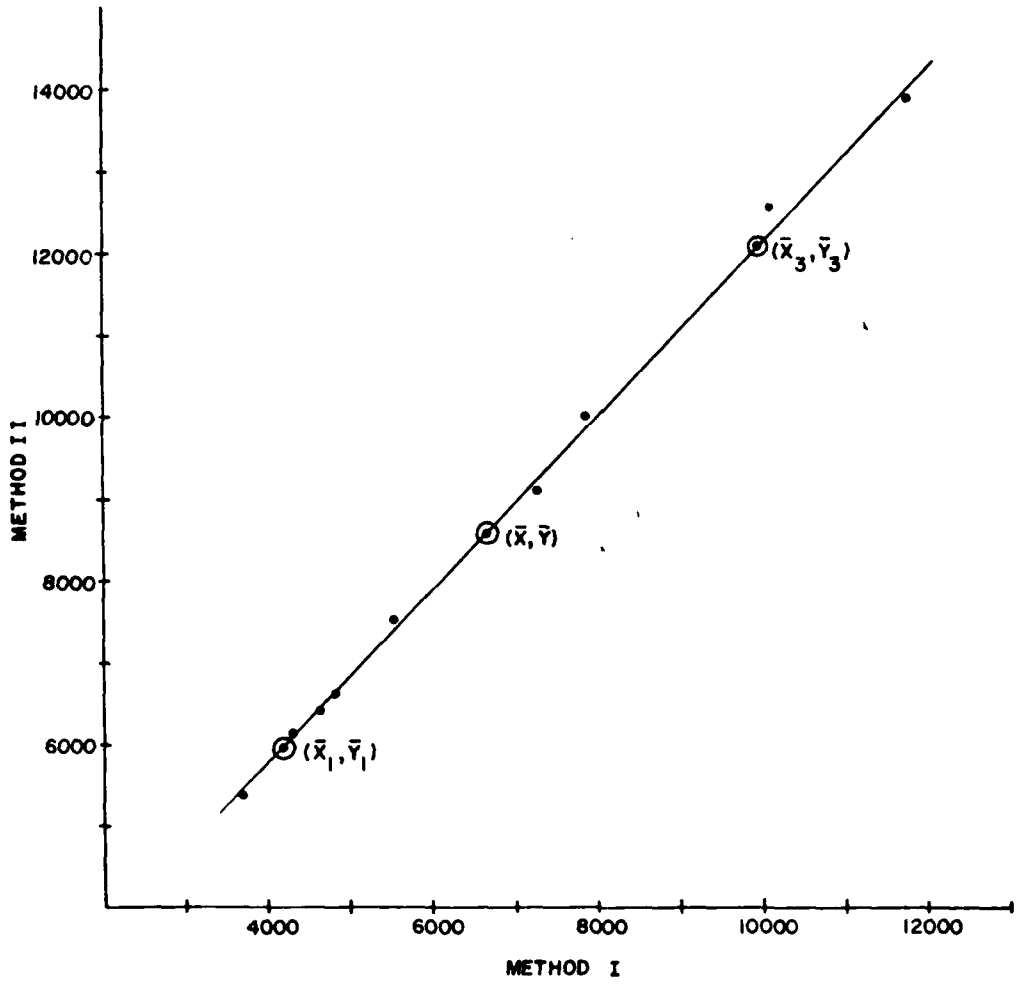


Figure 5-9. Relationship between two methods of determining a chemical constituent—an FII relationship.

similar to those appropriate to the least-squares method in FI situations, but involve more complex calculations. We do not consider them further here.

5-4.3.2 An Important Exceptional Case. Until comparatively recently it was not realized that there is a broad class of controlled experimental situations in which both X and Y are subject to errors of measurement, yet all of the techniques appropriate to the FI case (x's accurately known, measurement errors affect the Y's only) are strictly applicable without change.

As an example, let us consider the case of an analytical chemist who, in order to obtain an accurate determination of the concentration of a potassium sulphate solution, decides to proceed as follows: From a burette he will draw off 5, 10, 15, and 20 *ml* samples of the solution. Volume of solution is his independent variable x, and his target values are $x_1 = 5$, $x_2 = 10$, $x_3 = 15$, and $x_4 = 20$, respectively. The volumes of solution that he actually draws off X_1, X_2, X_3, and X_4 will, of course, differ from the nominal or target values as a result of errors of technique, and he will not attempt to measure their volumes accurately. These four samples of the potassium sulphate solution then will be treated with excess barium chloride, and the precipitated barium sulphate dried and weighed. Let Y_1, Y_2, Y_3, and Y_4 denote the corresponding yields of barium sulphate. These yields actually will correspond, of course, to the actual inputs X_1, X_2, X_3, and X_4, respectively; and will differ from the true yields associated with these inputs, say $y_1(X_1)$, $y_2(X_2)$, $y_3(X_3)$, and $y_4(X_4)$, respectively, as a result of errors of weighing and analytical technique. The sulphate concentration of the original potassium sulphate solution then will be determined by evaluating the slope b_1 of the best fitting straight line $Y = b_0 + b_1x$, relating the observed barium sulphate yields (Y_1, Y_2, Y_3, and Y_4) to the nominal or target volumes of solution (x_1, x_2, x_3, and x_4)—the intercept b_0 of the line making appropriate allowance for the possibility of bias of the analytical procedure resulting in a non-zero blank.

Without going into the merits of the foregoing as an analytical procedure, let us note a number of features that are common to *controlled experiments:* First, the experimental program involves a number of *preassigned* nominal or target values (x_1, x_2, ...) of the independent variable x, to which the experimenter equates the independent variable in his experiment as best he can, and then observes the corresponding yields (Y_1, Y_2, ...) of the dependent variable y; Second, the experimenter, in his notebook, records the *observed* yields (Y_1, Y_2, ...) as corresponding to, and treats them as if they were produced by, the *nominal* or *target* values (x_1, x_2, ...) of the independent variable—whereas, strictly they correspond to, and were produced by, the actual input values (X_1, X_2, ...), which ordinarily will differ somewhat from the nominal or target values (x_1, x_2, ...) as a result of errors of technique. Furthermore, the *effective values* (X_1, X_2, ...) of the independent variable actually realized in the experiment are not recorded at all—nor even measured!

It is surprising but nevertheless true that an underlying linear structural relationship of the form $y = \beta_0 + \beta_1x$ can be estimated validly from the results of such experiments, by fitting a line of the form $Y = b_0 + b_1x$ in accordance with the procedures for FI situations (x's known accurately, Y's only subject to error). This fact was emphatically brought to the attention of the scientific world by Joseph Berkson in a paper[6] published in 1950, and for its validity requires only the usual assumptions regarding the randomness and independence of the errors of measurement and technique affecting both of the variables (i.e., causing the deviations of the actual *inputs* X_1, X_2, ..., from their target values x_1, x_2, ..., *and* the deviations of the observed *outputs* Y_1, Y_2, ..., from their *true values* of $y_1(X_1)$, $y_2(X_2)$, ...). The conclusion also extends to the many-variable case considered in Chapter 6, *provided that* the relationship is linear, i.e., that

$$y = \beta_0 + \beta_1x + \beta_2u + \beta_3v + \ldots .$$

If the underlying relationship is a polynomial in x (e.g., $y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3$), then Geary[7] has found that Berkson's conclusion carries over to the extent that the usual least-squares estimates (given in Chapter 6) of the coefficients of the two highest powers of x (i.e., of β_2 and β_3 here) retain their optimum properties of unbiasedness and minimum variance, but

the confidence-interval and tests-of-significance procedures require modification.

5-4.4 SOME LINEARIZING TRANSFORMATIONS

If the form of a non-linear relationship between two variables is known, it is sometimes possible to make a transformation of one or both variables such that the relationship between the transformed variables can be expressed as a straight line. For example, we might know that the relationship is of the form $Y = ab^X$. If we take logs of both sides of this equation, we obtain

$$\log Y = \log a + X \log b,$$

which will be recognized to be a straight line whose intercept on the log Y scale is equal to log a, and whose slope is equal to log b. The procedure for fitting the relationship is given in the following steps.

(1) Make the transformation $Y_T = \log Y$ (i.e., take logs of all the observed Y values).

(2) Use the procedure of Paragraph 5-4.1.1 to fit the line $Y_T = b_0 + b_1X$, substituting Y_T everywhere for Y.

(3) Obtain the constants of the original equation by substituting the calculated values of b_0 and b_1 in the following equations:

$$b_0 = \log a$$
$$b_1 = \log b,$$

and taking the required antilogs.

Some relationships between X and Y which can easily be transformed into straight-line form are shown in Table 5-4. This table gives the appropriate change of variable for each relationship, and gives the formulas to convert the constants of the resulting straight line to the constants of the relationship in its original form. In addition to the ones given in Table 5-4, some more-complicated relationships can be handled by using special tricks which are not described here, but can be found in Lipka,[8] Rietz,[9] and Scarborough.[10]

It should be noted that the use of these transformations is certain to accomplish one thing only—i.e., to yield a relationship in straight-line form. The transformed data will not necessarily satisfy certain assumptions which are theoretically necessary in order to apply the procedures of Paragraph 5-4.1.1, for example, the assumption that the variability of Y given X is the same for all X. However, for practical purposes and within the range of the data considered, the transformations often do help in this regard.

Thus far, our discussion has centered on the use of transformations to convert a *known* relationship to linear form. The existence of such linearizing transformations also makes it possible to determine the form of a relationship empirically. The following possibilities, adapted from Scarborough,[10] are suggested in this regard:

(1) Plot Y against $\frac{1}{X}$ on ordinary graph paper. If the points lie on a straight line, the relationship is

$$Y = a + \frac{b}{X} .$$

(2) Plot $\frac{1}{Y}$ against X on ordinary graph paper. If the points lie on a straight line, the relationship is

$$Y = \frac{1}{a + bX}, \quad \text{or}$$
$$\frac{1}{Y} = a + bX .$$

(3) Plot X against Y on semilog paper (X on the arithmetic scale, Y on the logarithmic scale). If the points lie on a straight line, the variables are related in the form

$$Y = ae^{bX}, \text{ or}$$
$$Y = ab^X .$$

(4) Plot Y against X on log-log paper. If the points lie on a straight line, the variables are related in the form

$$Y = aX^b .$$

TABLE 5-4. SOME LINEARIZING TRANSFORMATIONS

If the Relationship Is of the Form:	Plot the Transformed Variables		Fit the Straight Line $Y_T = b_0 + b_1 X_T$	Convert Straight Line Constants (b_0 and b_1) To Original Constants:	
	$Y_T =$	$X_T =$		$b_0 =$	$b_1 =$
$Y = a + \frac{b}{X}$	Y	$\frac{1}{X}$	Use the procedures of Paragraph 5-4.1.1. In all formulas given there, substitute values of Y_T for Y and values of X_T for X, as appropriate.	a	b
$Y = \frac{1}{a + bX}$, or $\frac{1}{Y} = a + bX$	$\frac{1}{Y}$	X		a	b
$Y = \frac{X}{a + bX}$	$\frac{X}{Y}$	X		a	b
$Y = ab^X$	$\log Y$	X		$\log a$	$\log b$
$Y = ae^{bX}$	$\log Y$	X		$\log a$	$b \log e$
$Y = aX^b$	$\log Y$	$\log X$		$\log a$	b
$Y = a + bX^n$, where n is known	Y	X^n		a	b

5-5 PROBLEMS AND PROCEDURES FOR STATISTICAL RELATIONSHIPS

5-5.1 SI RELATIONSHIPS

In this case, we are interested in an association between two variables. See Paragraph 5-3.2 and Table 5-1.

We usually make the assumption that for any fixed value of X, the corresponding values of Y form a normal distribution with means $\bar{Y}_X = \beta_0 + \beta_1 X$ and variance $\sigma^2_{Y \cdot X}$ (read as "variance of Y given X") which is constant for all values of X.* Similarly, we usually assume that for any fixed value of Y, the corresponding values of X form a normal distribution with mean $\bar{X}_Y = \beta'_0 + \beta'_1 Y$ and variance $\sigma^2_{X \cdot Y}$, (variance of X given Y) which is constant for all values of Y.* Taken together, these two sets of assumptions imply that X and Y are jointly distributed according to the bivariate normal distribution. In practical situations, we usually have only a sample from all the possible pairs of values X and Y, and therefore we cannot determine either of the *true* regression lines, $\bar{Y}_X = \beta_0 + \beta_1 X$ or $\bar{X}_Y = \beta'_0 + \beta'_1 Y$, exactly. If we have a random sample of n pairs of values $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, we can estimate either line, or both. Our method of fitting the line gives us best predictions in the sense that, for a given $X = X'$ our estimate of the corresponding value of $Y = Y'$ will:

(a) on the average equal $\bar{Y}_{X'}$ the mean value of Y for $X = X'$ (i.e., it will be on the *true* line $\bar{Y}_X = \beta_0 + \beta_1 X$); and

(b) have a smaller variance than had we used any other method for fitting the line.

* Strictly, we should write

$m_{Y \cdot X} = \beta_0 + \beta_1 X$

and

$m_{X \cdot Y} = \beta'_0 + \beta'_1 Y$.

See Footnote in Paragraph 5-3.2.

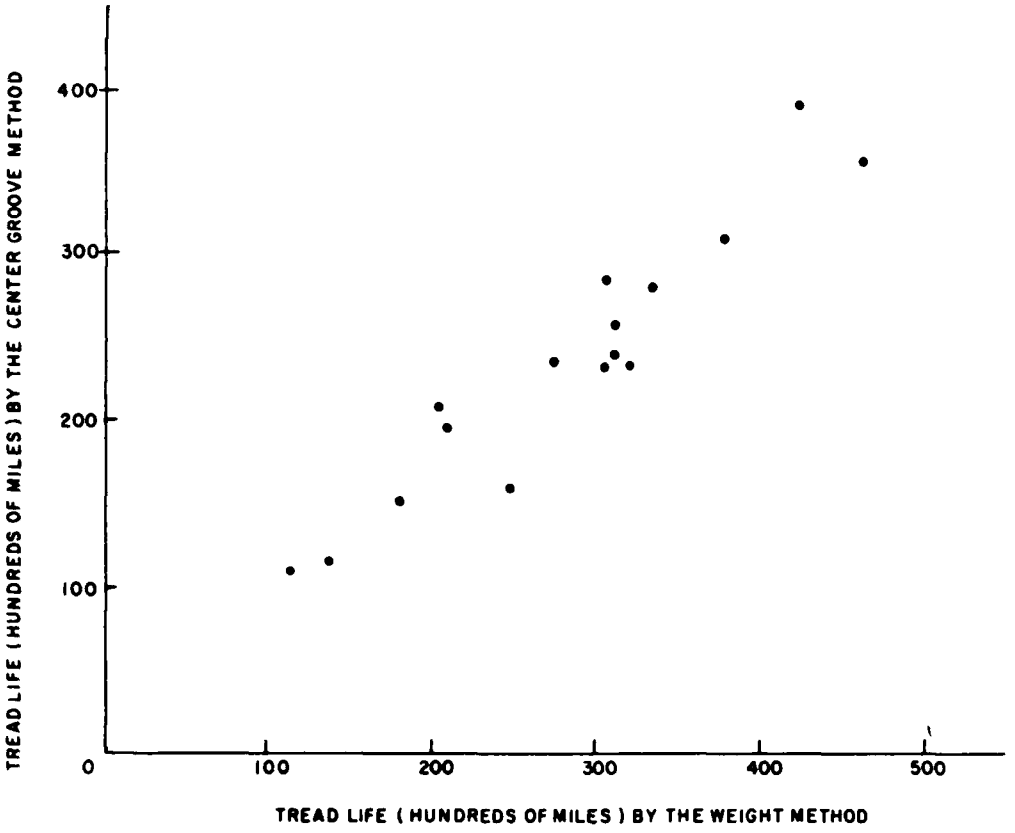


Figure 5-10. Relationship between the weight method and the center groove method of estimating tread life— an SI relationship.

Data Sample 5-5.1—Estimated Tread Wear of Tires

The data used for illustration are from a study of two methods of estimating tread wear of commercial tires (Stiehler and others[11]). The data are shown here and plotted in Figure 5-10. The variable which is taken as the independent variable X is the estimated tread life in hundreds of miles by the *weight-loss* method. The associated variable Y is the estimated tread life by the *groove-depth* method (center grooves). The plot seems to indicate a relationship between X and Y, but the relationship is statistical rather than functional or exact. The scatter of the points stems primarily from product variability and variation of tread wear under normal operating conditions, rather than from errors of measurement of weight loss or groove depth. Descriptions and predictions are applicable only "on the average."

X = Tread Life (Hundreds of Miles) Estimated By Weight Method	Y = Tread Life (Hundreds of Miles) Estimated By Center Groove Method
459	357
419	392
375	311
334	281
310	240
305	287
309	259
319	233
304	231
273	237
204	209
245	161
209	199
189	152
137	115
114	112

5-5.1.1 What is the Best Line To Be Used for Estimating $\bar{Y}_X$ for Given Values of X?

Procedure

The procedure is identical to that of Paragraph 5-4.1.1. Using Basic Worksheet (see Worksheet 5-5.1), compute the line

$$Y = b_0 + b_1 X.$$

This is an estimate of the true regression line

$$\bar{Y}_X = \beta_0 + \beta_1 X.$$

Using Data Sample 5-5.1, the equation of the fitted line is

$$Y = 13.506 + 0.790212\, X.$$

In Figure 5-11, the line is drawn, and confidence limits for the line (see Paragraph 5-5.1.2) are shown.

WORKSHEET 5-5.1

EXAMPLE OF SI RELATIONSHIP

X denotes Tread Life Estimated by Weight Method	Y denotes Tread Life Estimated by Center Groove Method
ΣX = 4505	ΣY = 3776
$\bar{X}$ = 281.5625	$\bar{Y}$ = 236

Number of points: n = 16

Step (1) ΣXY = 1,170,731

(2) $(\Sigma X)(\Sigma Y)/n$ = 1,063,180

(3) S_{xy} = 107551

(4) ΣX^2	1,404,543	(7) ΣY^2	985740
(5) $(\Sigma X)^2/n$	1,268,439.0625	(8) $(\Sigma Y)^2/n$	891136
(6) S_{xx}	136103.9375	(9) S_{yy}	94604

(10) $b_1 = \frac{S_{xy}}{S_{xx}}$	.790212	(14) $\frac{(S_{xy})^2}{S_{xx}}$	84988.119
(11) $\bar{Y}$	236	(15) $(n-2)\,s_Y^2$	9615.881
(12) $b_1\bar{X}$	222.494	(16) s_Y^2	686.849
(13) $b_0 = \bar{Y} - b_1\bar{X}$	13.506	s_Y	26.21

Equation of the line:

$Y = b_0 + b_1X$

$= 13.506 + .790212\,X$

s_{b_1} = 0.0710387

s_{b_0} = 21.048

Estimated variance of the slope:

$s_{b_1}^2 = \frac{s_Y^2}{S_{xx}}$ = .005046504

Estimated variance of intercept:

$s_{b_0}^2 = s_Y^2\left\{\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right\}$ = 443.002

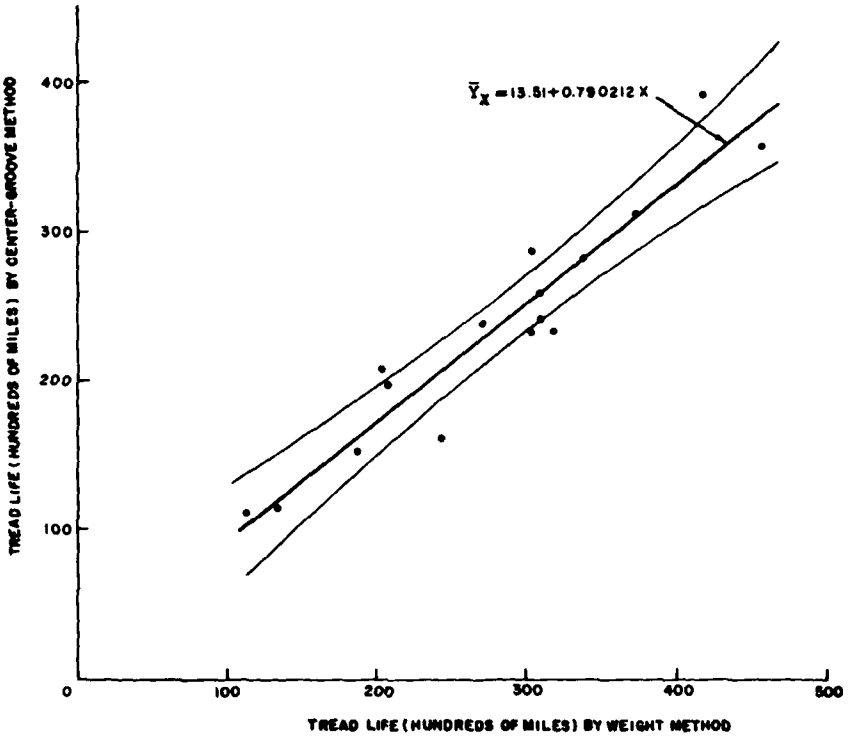


Figure 5-11. Relationship between weight method and center groove method—the line shown with its confidence band is for estimating tread life by center groove method from tread life by weight method.

Using the Regression Line for Prediction. The equation of the fitted line may be used to predict $\bar{Y}_X$, the average value of Y associated with a value of X. For example, using the fitted line, $Y = 13.506 + 0.790212\,X$, the following are some predicted values for $\bar{Y}_X$.

X	$\bar{Y}_X$
200	172
250	211
300	251
350	290
400	330
450	369

5-5.1.2 What are the Confidence Interval Estimates for: the Line as a Whole; a Point on the Line; a Single Y Corresponding to a New Value of X?

Read the discussion of the interpretation of three types of confidence intervals in Paragraph 5-4.1.2, in order to decide which is the appropriate kind of confidence interval.

The solutions are identical to those given in Paragraph 5-4.1.2, and are illustrated for the tread wear of commercial tires example (Data Sample 5-5.1).

5-5.1.2.1 What Is the $(1 - \alpha)$ Confidence Band for the Line as a Whole?

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Obtain s_Y from Worksheet.	(2) $s_Y = 26.21$
(3) Look up $F_{1-\alpha}$ for $(2, n - 2)$ degrees of freedom in Table A-5.	(3) $n = 16$ $F_{.95}(2, 14) = 3.74$
(4) Choose a number of values of X (within the range of the data) at which to compute points for drawing the confidence band.	(4) Let: $X = 200$ $X = 250$ $X = 300$ $X = 350$ $X = 400$, for example.
(5) At each selected value of X, compute: $Y_c = \bar{Y} + b_1 (X - \bar{X})$ and $W_1 = \sqrt{2F}\, s_Y \left[\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{xx}}\right]$	(5) See Table 5-5 for a convenient computational arrangement, and the example calculations.
(6) A $(1 - \alpha)$ confidence band for the whole line is determined by $Y_c \pm W_1$.	(6) See Table 5-5.
(7) To draw the line and its confidence band, plot Y_c at two of the extreme selected values of X. Connect the two points by a straight line. At each selected value of X, plot also $Y_c + W_1$ and $Y_c - W_1$. Connect the upper series of points, and the lower series of points, by smooth curves.	(7) See Figure 5-11.
If more points are needed for drawing the curves, note that, because of symmetry, the calculation of W_1 at n values of X actually gives W_1 at $2n$ values of X.	For example: W_1 (but not Y_c) has the same value at $X = 250$ (i.e., $\bar{X} - 31.56$) as at $X = 313.12$ (i.e., $\bar{X} + 31.56$).

TABLE 5-5. COMPUTATIONAL ARRANGEMENT FOR PROCEDURE 5-5.1.2.1

X	$(X - \bar{X})$	Y_c	$\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{xx}}$	$s^2_{Y_c}$	s_{Y_c}	W_1	$Y_c + W_1$	$Y_c - W_1$
200	−81.56	171.6	0.111375	76.50	8.746	23.9	195.5	147.7
250	−31.56	211.1	0.069818	47.95	6.925	18.9	230.0	192.2
300	+18.44	250.6	0.064998	44.64	6.681	18.3	268.9	232 3
350	68.44	290.1	0.096915	66.57	8.159	22.3	312.4	267.8
400	118.44	329.6	0.165569	113.72	10.66	29.2	358.8	300.4

$\bar{X} = 281.5625$
$\bar{Y} = 236$

$s^2_Y = 686.849$
$\frac{1}{n} = .0625$
$b_1 = 0.790212$
$S_{xx} = 136103.9375$

$Y_c = \bar{Y} + b_1 (X - \bar{X})$

$s^2_{Y_c} = s^2_Y \left[\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{xx}}\right]$

$\sqrt{2F} = \sqrt{7.48}$
$= 2.735$

$W_1 = \sqrt{2F}\, s_{Y_c}$

5-5.1.2.2 Give a $(1 - \alpha)$ Confidence Interval Estimate For a Single Point On the Line, i.e., the Mean Value of Y Corresponding to $X = X'$.

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Obtain s_Y from Worksheet.	(2) $s_Y = 26.21$
(3) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(3) $n = 16$ $t_{.975}$ for 14 d.f. $= 2.145$
(4) Choose X', the value of X at which we want to make an interval estimate of the mean value of Y.	(4) Let $X' = 250$, for example.
(5) Compute: $W_2 = t_{1-\alpha/2}\, s_Y \left[\frac{1}{n} + \frac{(X' - \bar{X})^2}{S_{xx}}\right]^{\frac{1}{2}}$ and $Y_c = \bar{Y} + b_1 (X' - \bar{X})$	(5) $W_2 = (2.145)\ (26.21)\ (.2642)$ $= 14.85$ $Y_c = 211.1$
(6) A $(1 - \alpha)$ confidence interval estimate for the mean value of Y corresponding to $X = X'$ is given by $Y_c \pm W_2$.	(6) A 95% confidence interval estimate for the mean value of Y corresponding to $X = 250$ is 211.1 ± 14.8 , the interval from 196.3 to 225.9 .

5-5.1.2.3 Give a $(1 - \alpha)$ Confidence Interval Estimate For a Single (Future) Value of Y Corresponding to a Chosen Value of $X = X'$.

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Obtain s_Y from Worksheet.	(2) $s_Y = 26.21$
(3) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(3) $n = 16$ $t_{.975}$ for 14 d.f. $= 2.145$
(4) Choose X', the value of X at which we want to make an interval estimate of a single value of Y.	(4) Let $X' = 250$, for example.
(5) Compute: $W_3 = t_{1-\alpha/2}\, s_Y \left[1 + \frac{1}{n} + \frac{(X' - \bar{X})^2}{S_{xx}}\right]^{\frac{1}{2}}$ and $Y_c = \bar{Y} + b_1 (X' - \bar{X})$	(5) $W_3 = (2.145)\ (26.21)\ (1.0343)$ $= 58.1$ $Y_c = 211.1$.
(6) A $(1 - \alpha)$ confidence interval estimate for Y' (the single value of Y corresponding to X') is $Y_c \pm W_3$.	(6) A 95% confidence interval estimate for a single value of Y corresponding to $X' = 250$ is 211.1 ± 58.1, the interval from 153.0 to 269.2 .

5-5.1.3 Give a Confidence Interval Estimate For β_1, the Slope of the True Regression Line, $\bar{Y}_X = \beta_0 + \beta_1 X$.

The solution is identical to that of Paragraph 5-4.1.3 and is illustrated here for Data Sample 5-5.1 .

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(2) $n = 16$ $t_{.975}$ for 14 d.f. $= 2.145$
(3) Obtain s_{b_1} from Worksheet.	(3) $s_{b_1} = 0.0710387$
(4) Compute $W_4 = t_{1-\alpha/2}\, s_{b_1}$	(4) $W_4 = (2.145)\ (.0710387)$ $= 0.152378$
(5) A $(1 - \alpha)$ confidence interval estimate for β_1 is $b_1 \pm W_4$.	(5) $b_1 = 0.790212$ $W_4 = 0.152378$ A 95% confidence interval estimate for β_1 is the interval 0.790212 ± 0.152378, i.e., the interval from 0.637834 to 0.942590 .

5-5.1.4 What Is the Best Line For Predicting $\bar{X}_Y$ From Given Values of Y?

For this problem, we fit a line $X = b_0' + b_1' Y$ (an estimate of the true line $\bar{X}_Y = \beta_0' + \beta_1' Y$). To fit this line we need to interchange the roles of the X and Y variables in the computations outlined in Worksheet 5-5.1 and proceed as in Paragraph 5-5.1.1.

That is, the fitted line will be:

$$X = b_0' + b_1' Y \ ,$$

where

$$b_0' = \bar{X} - b_1' \bar{Y}$$

and

$$b_1' = \frac{S_{xy}}{S_{yy}} \ .$$

From Data Sample 5-5.1:

$$b_1' = \frac{107551}{94604} = 1.136855$$

$$b_0' = 281.5625 - (1.136855)\ (236) = 13.26$$

The equation of the fitted line is:

$$X = 13.26 + 1.136855\ Y \ ,$$

and this line is shown in Figure 5-12, along with the line for predicting Y from X.

In order to obtain confidence intervals, we need the following formulas:

$$s_X^2 = \frac{S_{xx} - \frac{(S_{xy})^2}{S_{yy}}}{n-2}$$

$$s_{b_1'}^2 = \frac{s_X^2}{S_{yy}}$$

$$s_{b_0'}^2 = s_X^2 \left\{ \frac{1}{n} + \frac{(\bar{Y})^2}{S_{yy}} \right\} \ .$$

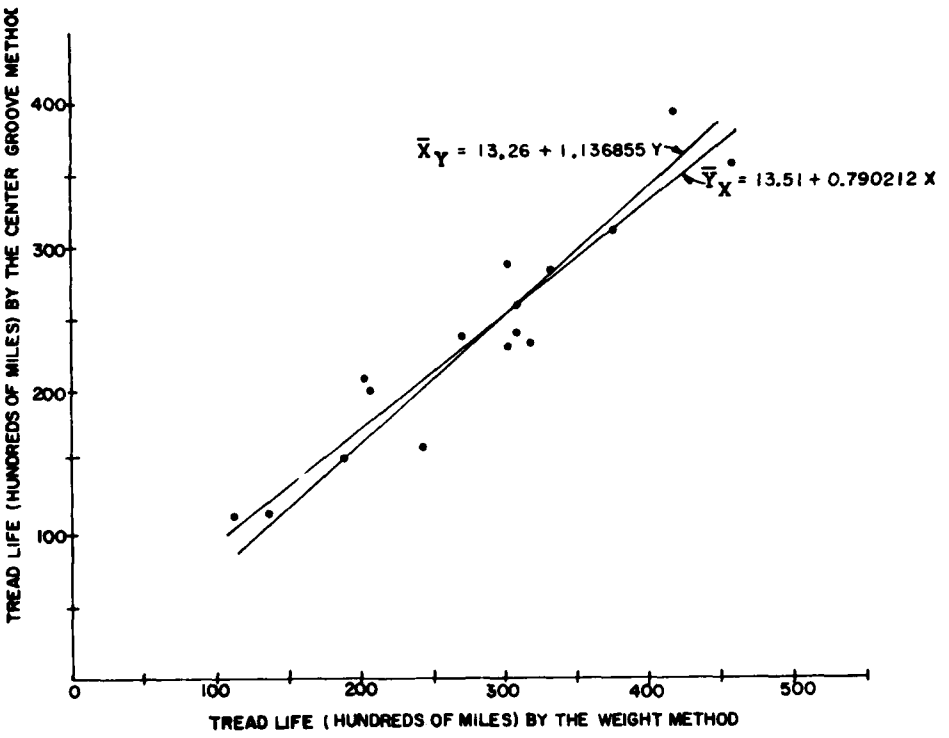


Figure 5-12. Relationship between weight method and center groove method—showing the two regression lines.

5-5.1.5 What is the Degree of Relationship of the Two Variables X and Y as Measured by ρ, the Correlation Coefficient?

Procedure

(1) Compute

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

(2) A 95% confidence interval for ρ can be obtained from Table A-17, using the appropriate n and r. If the confidence interval does not include $\rho = 0$, we may state that the data give reason to believe that there is a relationship (measured by $\rho \neq 0$) between the two variables; otherwise, we may state that the data are consistent with the possibility that the two variables are uncorrelated ($\rho = 0$).

Example

(1) Using Worksheet 5-5.1,

$$r = \frac{107551}{\sqrt{136103.94}\sqrt{94604}}$$

$$= \frac{107551}{(368.92)(307.58)}$$

$$= 0.95 \ .$$

(2) $n = 16$
$r = 0.95$

From Table A-17, the 95% confidence interval estimate of ρ is the interval from 0.85 to 0.98. Since this interval does not include $\rho = 0$, we may state that the data give reason to believe that there is a relationship between the two methods of estimating tread wear of tires.

5-5.2 SII RELATIONSHIPS

In this case, we are interested in an association between two variables. This case differs from SI in that one variable has been measured at only preselected values of the other variable. (See Paragraph 5-3.2 and Table 5-1.)

For any given value of X, the corresponding values of Y have a normal distribution with mean $\bar{Y}_X = \beta_0 + \beta_1 X$, and variance $\sigma^2_{Y \cdot X}$ which is independent of the value of X. We have n pairs of values $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, in which X is the independent variable. (The X values are selected, and the Y values are thereby determined.) We wish to describe the line which will enable us to make the best estimate of values of Y corresponding to given values of X.

We have seen that for SI there are two lines, one for predicting Y from X and one for predicting X from Y. When we use only selected values of X, however, the only appropriate line to fit is $Y = b_0 + b_1 X$.

It should be noted that SII is handled computationally in the same manner as FI, but both the underlying assumptions and the interpretation of the end results are different.

Data Sample 5-5.2—Estimated Tread Wear of Tires

For our example, we use part of the data used in Data Sample 5-5.1 (the SI example). Suppose that, due to some limitation, we were only able to measure X values between $X = 200$ and $X = 400$, or that we had taken but had lost the data for $X < 200$ and $X > 400$. From Figure 5-10, we use only the 11 observations whose X values are between these limits. The "selected" data are recorded in the following table.

X = Tread Life (Hundreds of Miles) Estimated By Weight Method	Y = Tread Life (Hundreds of Miles) Estimated By Center Groove Method
375	311
334	281
310	240
305	287
309	259
319	233
304	231
273	237
204	209
245	161
209	199

5-5.2.1 What Is the Best Line To Be Used for Estimating $\bar{Y}_X$ From Given Values of X?

Procedure

Using Basic Worksheet (see Worksheet 5-5.2), compute the line $Y = b_0 + b_1X$. This is an estimate of the true line $\bar{Y}_X = \beta_0 + \beta_1X$.

Using Data Sample 5-5.2, the fitted line is

$$Y = 48.965 + 0.661873\ X.$$

The fitted line is shown in Figure 5-13, and the confidence band for the line (see the procedure of Paragraph 5-5.2.2.1) also is shown.

WORKSHEET 5-5.2

EXAMPLE OF SII RELATIONSHIP

X denotes Tread Life Estimated by Weight Method

Y denotes Tread Life Estimated by Center Groove Method

ΣX =	3187	ΣY =	2648
$\bar{X}$ =	289.727	$\bar{Y}$ =	240.727

Number of points: n = 11

Step (1)	ΣXY	=	785369		
(2)	$(\Sigma X)(\Sigma Y)/n$	=	767197.818		
(3)	S_{xy}	=	18171.182		
(4)	ΣX^2	=	950815	(7) ΣY^2	= 655754
(5)	$(\Sigma X)^2/n$	=	923360.818	(8) $(\Sigma Y)^2/n$	= 637445.818
(6)	S_{xx}	=	27454.182	(9) S_{yy}	= 18308.182
(10)	$b_1 = \frac{S_{xy}}{S_{xx}}$	=	0.661873	(14) $\frac{(S_{xy})^2}{S_{xx}}$	= 12027.015
(11)	$\bar{Y}$	=	240.727	(15) $(n-2)\,s_Y^2$	= 6281.167
(12)	$b_1\bar{X}$	=	191.762	(16) s_Y^2	= 697.9074
(13)	$b_0 = \bar{Y} - b_1\bar{X}$	=	48.965	s_Y	= 26.418

Equation of the line:

$Y = b_0 + b_1X$

$= 48.965 + 0.661873\ X$

s_{b_1} = 0.159439

s_{b_0} = 46.88

Estimated variance of the slope:

$s_{b_1}^2 = \frac{s_Y^2}{S_{xx}}$ = .0254208

Estimated variance of intercept:

$s_{b_0}^2 = s_Y^2\left\{\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right\}$ = 2197.313

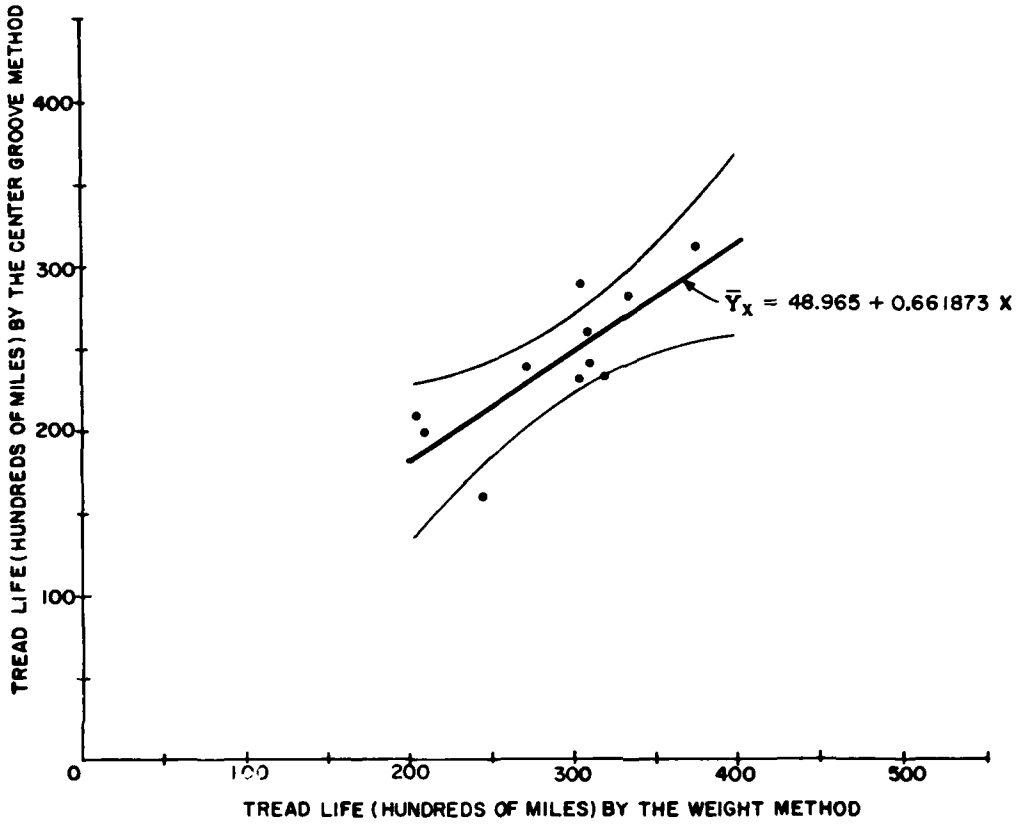


Figure 5-13. Relationship between weight method and center groove method when the range of the weight method has been restricted—an SII relationship.

5-5.2.2 What are the Confidence Interval Estimates for: the Line as a Whole; a Point on the Line; a Single Y Corresponding to a New Value of X?

Read the discussion of the interpretation of these three types of confidence intervals in Paragraph 5-4.1.2 in order to decide which is the appropriate kind of confidence interval.

5-5.2.2.1 What Is the $(1 - \alpha)$ Confidence Band For the Line as a Whole?

The solution is identical to that of Procedure 5-4.1.2.1 and is illustrated here for Data Sample 5-5.2.

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Obtain s_Y from Worksheet.	(2) From Worksheet 5-5.2 $s_Y = 26.418$
(3) Look up $F_{1-\alpha}$ for $(2, n - 2)$ degrees of freedom in Table A-5.	(3) $n = 11$ $F_{.95}(2, 9) = 4.26$
(4) Choose a number of values of X (within the range of the data) at which to compute points for drawing the confidence band.	(4) Let: $X = 200$ $X = 250$ $X = 300$ $X = 350$ $X = 400$, for example.
(5) At each selected value of X, compute: $Y_c = \bar{Y} + b_1 (X - \bar{X})$ and $W_1 = \sqrt{2F}\, s_Y \left[\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{xx}}\right]^{\frac{1}{2}}$	(5) See Table 5-6 for a convenient computational arrangement and the example calculations.
(6) A $(1 - \alpha)$ confidence band for the whole line is determined by $Y_c \pm W_1$.	(6) See Table 5-6.
(7) To draw the line and its confidence band, plot Y_c at two of the extreme selected values of X. Connect the two points by a straight line. At each selected value of X, also plot $Y_c + W_1$ and $Y_c - W_1$. Connect the upper series of points, and the lower series of points, by smooth curves.	(7) See Figure 5-13.
If more points are needed for drawing the curves for the band, note that, because of symmetry the calculation of W_1 at n values of X actually gives W_1 at $2n$ values of X.	For example: W_1 (but not Y_c) has the same value at $X = 250$ (i.e., $\bar{X} - 39.73$) as at $X = 329.5$ (i.e., $\bar{X} + 39.73$).

TABLE 5-6. COMPUTATIONAL ARRANGEMENT FOR PROCEDURE 5-5.2.2.1

X	$(X - \bar{X})$	Y_c	$\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{xx}}$	$s^2_{Y_c}$	s_{Y_c}	W_1	$Y_c + W_1$	$Y_c - W_1$
200	−89.73	181.3	0.384179	268.12	16.37	47.8	229.1	133.5
250	−39.73	214.4	0.148404	103.57	10.18	29.7	244.1	184.7
300	+10.27	247.5	0.094751	66.127	8.132	23.7	271.2	223.8
350	60.27	280.6	0.223219	155.79	12.48	36.4	317.0	244.2
400	110.27	313.7	0.533810	372.55	19.30	56.3	370.0	257.4

$\bar{X} = 289.727$
$\bar{Y} = 240.727$

$s^2_Y = 697.9074$
$\frac{1}{n} = 0.0909091$
$b_1 = 0.661873$
$S_{xx} = 27454.182$

$Y_c = \bar{Y} + b_1 (X - \bar{X})$

$s^2_{Y_c} = s^2_Y \left[\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{xx}}\right]$

$\sqrt{2F} = \sqrt{8.52} = 2.919$

$W_1 = \sqrt{2F}\, s_{Y_c}$

5-5.2.2.2 Give a $(1 - \alpha)$ Confidence Interval For a Single Point On the Line, i.e., the Mean Value of Y Corresponding To a Chosen Value of X (X').

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Obtain s_Y from Basic Worksheet.	(2) From Worksheet 5-5.2 $s_Y = 26.418$
(3) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(3) $n = 11$ $t_{.975}$ for 9 d.f. $= 2.262$
(4) Choose X', the value of X at which we want to make an interval estimate of the mean value of Y.	(4) Let $X' = 300$, for example.
(5) Compute: $W_2 = t_{1-\alpha/2}\, s_Y \left[\frac{1}{n} + \frac{(X' - \bar{X})^2}{S_{xx}}\right]^{\frac{1}{2}}$ and $Y_c = \bar{Y} + b_1 (X' - \bar{X})$	(5) $W_2 = (2.262)\ (26.418)\ (0.3078)$ $= 18.4$ $Y_c = 247.5$
(6) A $(1 - \alpha)$ confidence interval estimate for the mean value of Y corresponding to $X = X'$ is given by $\bar{Y} + b_1 (X - \bar{X}) \pm W_2$ $= Y_c \pm W_2$.	(6) A 95% confidence interval estimate for the mean value of Y at $X = 300$ is the interval 247.5 ± 18.4, i.e., the interval from 229.1 to 265.9 .

5-5.2.2.3 Give a $(1 - \alpha)$ Confidence Interval Estimate For a Single (Future) Value of Y Corresponding To a Chosen Value of $X = X'$.

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Obtain s_Y from Worksheet.	(2) From Worksheet 5-5.2 $s_Y = 26.418$
(3) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(3) $t_{.975}$ for 9 d.f. $= 2.262$
(4) Choose X', the value of X at which we want to make an interval estimate of a single value of Y.	(4) Let $X' = 300$, for example.
(5) Compute: $W_3 = t_{1-\alpha/2}\, s_Y \left[1 + \frac{1}{n} + \frac{(X' - \bar{X})^2}{S_{xx}}\right]^{\frac{1}{2}}$ and $Y_c = \bar{Y} + b_1 (X' - \bar{X})$	(5) $W_3 = (2.262)\ (26.418)\ (1.0463)$ $= 62.5$ $Y_c = 247.5$
(6) A $(1 - \alpha)$ confidence interval estimate for Y' (the single value of Y corresponding to X') is given by $\bar{Y} + b_1 (X' - \bar{X}) \pm W_3$ $= Y_c \pm W_3$.	(6) A 95% confidence interval estimate for Y at $X = 300$ is the interval 247.5 ± 62.5, i.e., the interval from 185.0 to 310.0 .

5-5.2.3 What Is the Confidence Interval Estimate for β_1, the Slope of the True Line, $\bar{Y}_X = \beta_0 + \beta_1 X$?

Procedure	Example
(1) Choose the desired confidence level, $1 - \alpha$	(1) Let: $1 - \alpha = .95$ $\alpha = .05$
(2) Look up $t_{1-\alpha/2}$ for $n - 2$ degrees of freedom in Table A-4.	(2) $n = 11$ $t_{.975}$ for 9 d.f. $= 2.262$
(3) Obtain s_{b_1} from Worksheet.	(3) From Worksheet 5-5.2 $s_{b_1} = 0.159439$
(4) Compute $W_4 = t_{1-\alpha/2}\, s_{b_1}$	(4) $W_4 = 2.262\ (0.159439)$ $= 0.360651$
(5) A $(1 - \alpha)$ confidence interval estimate for β_1 is $b_1 \pm W_4$.	(5) $b_1 = 0.661873$ $W_4 = 0.360651$ A 95% confidence interval estimate for β_1 is the interval 0.661873 ± 0.360651, i.e., the interval from 0.301222 to 1.022524 .

REFERENCES

1. C. Eisenhart, "The Interpretation of Certain Regression Methods and Their Use in Biological and Industrial Research," *Annals of Mathematical Statistics*, Vol. 10, No. 2, pp. 162-186, June, 1939.
2. M. Ezekiel, *Methods of Correlation Analysis* (2d edition), Chapter 20, John Wiley & Sons, Inc., New York, N.Y., 1941.
3. J. Mandel, "Fitting a Straight Line to Certain Types of Cumulative Data," *Journal of the American Statistical Association*, Vol. 52, pp. 552-566, 1957.
4. F. S. Acton, *Analysis of Straight-Line Data*, John Wiley & Sons, Inc., New York, N.Y., 1959.
5. M. S. Bartlett, "Fitting a Straight Line When Both Variables are Subject to Error," *Biometrics*, Vol. 5, No. 3, pp. 207-212, 1949.
6. J. Berkson, "Are There Two Regressions?", *Journal of the American Statistical Association*, Vol. 45, pp. 164-180, 1950.
7. R. C. Geary, "Non-linear Functional Relationships Between Two Variables When One is Controlled," *Journal of the American Statistical Association*, Vol. 48, pp. 94-103, 1953.
8. J. Lipka, *Graphical and Mechanical Computation*, John Wiley & Sons, Inc., New York, N.Y., 1918.
9. H. L. Rietz, (ed.), *Handbook of Mathematical Statistics*, Houghton Mifflin Company, Boston, Mass., 1924.
10. J. B. Scarborough, *Numerical Mathematical Analysis* (2d edition), The Johns Hopkins Press, Baltimore, Md., 1950.
11. R. D. Stiehler, G. G. Richey, and J. Mandel, "Measurement of Treadwear of Commercial Tires," *Rubber Age*, Vol. 73, No. 2, May, 1953.

Reprinted from **ANALYTICAL CHEMISTRY**, Vol. 29, Page 743, May 1957

Study of Accuracy in Chemical Analysis Using Linear Calibration Curves

JOHN MANDEL and FREDERIC J. LINNIG
National Bureau of Standards, Washington 25, D. C.

▶ **In situations characterized by linear calibration curves such as the relation between "found" and "added" in studies of accuracy in chemical analysis, the usual method for deriving confidence intervals for the slope and the intercept of the fitted straight line *may lead to erroneous conclusions.* The difficulty results from the interdependence of multiple conclusions drawn from the same data, especially when there is a strong correlation between the parameters involved. The method of joint confidence regions eliminates these difficulties and has *the further advantage of allowing for* the evaluation of the uncertainty of the calibration line as a whole, as well as of any values or functions of values derived from it.**

THE STUDY of an analytical procedure generally starts with determining satisfactory operating conditions. Once this has been done, the precision and accuracy of the method can be effectively studied by analyzing a series of prepared samples covering the range of concentrations over which the method is applicable. This procedure, which involves the statistical theory of fitting straight lines based on the method of least squares, has been described by Youden (*18, 19*) and applied by Linnig, Mandel, and Peterson (*9*) and by Lark (*8*). Essentially, the slope of the fitted straight line can be compared to a value based on stoichiometric or other theoretical considerations; the intercept, to "blank" determinations; and the "standard error of estimate," to a measure of precision obtained from replicate determinations. Thus, the method generally involves tests of significance of the slope and the intercept of a fitted straight line. Such tests can readily be carried out in accordance with classical theory (*1, 2, 12, 19*).

Lark (*8*) has pointed out that the tests of significance on slope and intercept can lead to erroneous conclusions, because these tests, when carried out independently of each other, ignore the strong correlation that exists between the estimated slope and intercept of a straight line obtained by least squares calculations. In Table I, the values labeled "found" differ from those denoted "added" merely by random fluctuations. Thus, the "true" relation between found and added is a straight line passing through the origin with a slope equal to unity. This line is denoted T in Figure 1. The line E, on the other hand, which is the least squares fit

of the equation $y = b + mx$ to these data, has, as the result of the random errors, an intercept different from zero and a slope different from unity. It is easily seen that if by the interplay of chance effects, the fitted line has a slope less than unity, it will tend to have a positive intercept and vice versa. Thus, if the error in the slope is negative, the error in the intercept will tend to be positive, and vice versa. The theory of least squares shows this to be generally true for any set of linear data, for which the average of the x values (values "added") is positive; in these cases, the errors of slope and intercept are always negatively correlated, regardless of the precision of the data.

Table I. Illustrative Data

"Added"	"Found"
15	25.4
30	26.8
45	43.6
60	62.8
75	82.5
90	84.0

In this paper a rational basis is provided for judging the reliability of slope, intercept, and any value derived from the calibration line. The concepts are presented in terms of the data obtained by the authors in the study of an analytical method for which interesting chemical interpretations were suggested for some of the statistical conclusions (*9*). However, reference will also be made to the data of Table I to illustrate situations where the high reproducibility of the analytical method just referred to would produce effects too small to be distinguished graphically. For greater continuity of presentation, all mathematical and computational matters are relegated to a later section.

ACCURACY IN CHEMICAL ANALYSIS

Linnig, Mandel, and Peterson (*9*) obtained the data given in Table II for the determination of fatty acid in rubber.

Table II. Determination of Fatty Acid in Rubber

	Fatty Acid, Mg.	
Titration	Added	Found
1	20.0	28.0
2	20.0	24.5
3	50.0	58.5
4	50.0	57.8
5	150.0	157.8
6	153.7	163.2
7	250.0	257.8
8	250.0	259.3
9	500.0	512.4
10	500.0	509.2
Solvent blank	0.0	7.40

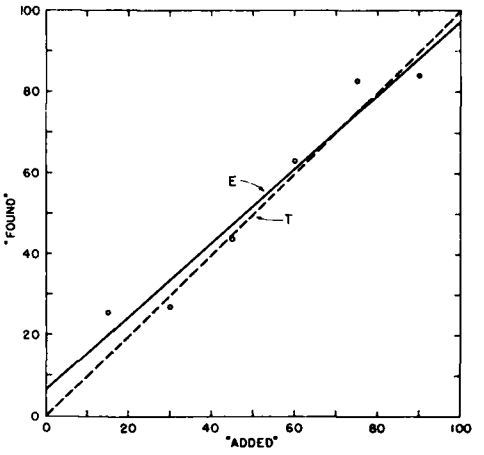


Figure 1. Effect of experimental errors on fitted straight line

Data of Table I
T. Theoretical line
E. Fitted line

From a chemical viewpoint the following questions are pertinent:

1. Does this analytical procedure require a blank correction?
2. Is the value for the blank that was determined experimentally (b = 7.40 mg.) an acceptable correction for the removal of the constant type of error suggested by the data?
3. Does the removal of the constant-type error (by means of a blank correction) lead to an otherwise accurate method? More specifically, is there, in addition to a constant-type error, also an error of a relative type—i.e., one that increases as the amount of material to be titrated increases?

These questions relate to the values of the intercept and the slope of a plot of "found" *vs.* "added," similar to the one shown in Figure 1 (*9*).

Now, if answers to these questions are obtained by means of a statistical analysis, these answers should be compatible with the data, not only individually but collectively. For example, it has been suggested (*9*) that the existence of a relative type of error (slope different from unity) in titration data of the type given in Table II is related to the choice of an indicator that does not change at the equivalence point. Therefore, in order to determine the adequacy of a particular indicator, one would test statistically the significance of the departure of the slope from unity. On the other hand, one may wish to judge the adequacy of a blank titration as a correction for a constant error by testing the significance of its difference from the observed intercept. Chemically, these may be entirely unrelated questions; but from the viewpoint of experimental evidence, they are related in that they

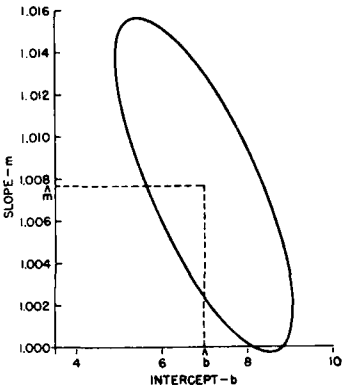


Figure 2. Joint confidence ellipse for slope and intercept

Data of Table II
$\hat{m}$. Least squares estimate of slope
$\hat{b}$. Least squares estimate of intercept

are obtained from the same set of data. This is especially so because, as has been indicated, there exists a strong statistical correlation between the errors in the slope and the intercept. Consequently, the answers to both questions must be jointly compatible with the data, and this requires the use of joint confidence regions.

JOINT RELIABILITY OF SLOPE AND INTERCEPT

A joint confidence region for slope and intercept is shown in Figure 2. On the abscissa point $\hat{b}$ represents the value for the intercept obtained by the method of least squares. Similarly, $\hat{m}$ on the

ordinate is the least squares estimate of the slope. Point ($\hat{b}$, $\hat{m}$) establishes, therefore, the line of "best fit." However, even this line of best fit is probably in error, the magnitude of its discrepancy from the true line depending on the experimental errors in the measurements to which the line was fitted. Consequently, points other than ($\hat{b}$, $\hat{m}$) are admissible, and theory shows (*12*, p. 296) that these points lie in an ellipse having the point of best fit as center. The boundary of the ellipse is determined by the magnitude of the experimental errors and by the degree of confidence, the "confidence coefficient," with which one wishes to state that the true point lies in the interior of the ellipse. The tilt of the ellipse with respect to the axes is a consequence of the negative correlation between the errors in slope and intercept. As a result of the tilt, the ellipse favors points with a higher slope and lower intercept than the best fit (upper left area) and points with a lower slope and higher intercept (lower right area); while points corresponding to lower slopes and lower intercepts (lower left) or to higher slopes and higher intercepts (upper right) tend to fall outside the admissible region.

is required is answered by determining whether the ellipse contains points for which $b = 0$. As all such points are on a vertical line at $b = 0$, they are well outside the ellipse, and it is at once apparent that $b = 0$ is unacceptable. Consequently, a blank is required to correct for a constant-type error.

2. Is the experimental blank an adequate correction for the constant-type error? To answer this question, draw a vertical line at $b = 7.40$ (the value of the experimental blank). This line intersects the ellipse and is, in fact, close to its center. Consequently, there is no reason to doubt the validity of this blank as a means of correcting for the constant-type error.

3. Is there a relative-type error? Answering this question is equivalent to deciding whether the value, $m = 1$, is acceptable. Consider the horizontal line, $m = 1$. The points on this line falling inside the ellipse are extremely close to the boundary of the ellipse. Therefore, the hypothesis. $m = 1$, is of doubtful validity, and there exists a strong likelihood that, in addition to an error of the constant type, there is a relative type of error. It has been suggested in relation to these data (*9*) that the reason for finding results higher than the stoichiometric values may be the opacity of the solution, causing the change of color of the indicator to be observed somewhat beyond the equivalence point.

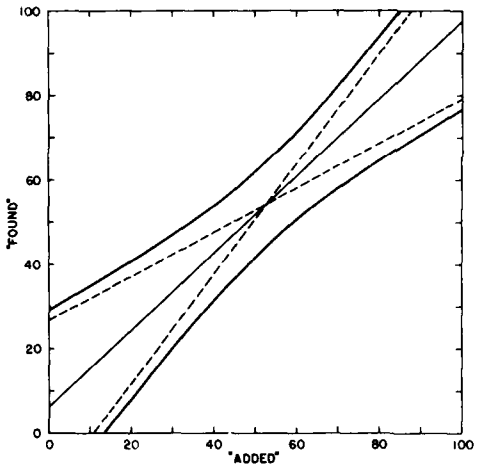


Figure 3. Confidence band for calibration line

Data of Table I. Straight line is least squares fit. The two branches of hyperbola define the confidence band. Broken lines are the asymptotes

APPLICATION TO CHEMICAL EXAMPLE

Figure 2 represents the 95% joint confidence ellipse corresponding to the data in Table II. The three questions that arose in connection with these data can readily be answered by means of this figure.

1. The question as to whether a blank

The selection of any particular value for the intercept, even though acceptable, restricts the range of acceptable slope values. Thus, as seen in Figure 2, if it is decided to use a blank correction of 7.40 mg., the range of slope values that are compatible with this blank extends approximately from 1.0013 to 1.0117—i.e., it is no longer equal to the total range enclosing the entire ellipse (approximately 1.000 to 1.016).

As the least-squares solution, $m = 1.00765$, is well within this restricted range, the procedure which consists of first correcting the data by means of the experimental blank and then dividing by 1.00765 is entirely acceptable. At the same time it is apparent that merely subtracting the experimental blank is not satisfactory, because this amounts to accepting the joint hypothesis, $b = 7.40$ and $m = 1$, which corresponds to a point outside the ellipse.

If the acceptability of the blank had been judged on the basis of a confidence interval obtained by the usual method—i.e., not based on the joint confidence region—then this judgment would, in a sense, have exhausted the confidence coefficient. If now a judgment concerning the true value of the slope were also attempted, then the joint judgment, concerning both intercept and slope, would no longer be associated with the initially chosen confidence coefficient. In view of the strong correlation between slope and intercept, any proposed value for the intercept restricts the choice of acceptable values for the slope and vice versa. This fact is ignored in the usual method of examining slope and intercept separately.

RELIABILITY OF CALIBRATION LINE

The study of an analytical procedure by the method of linear regression leads to values for the slope and the intercept of the calibration line, and by the method described in the preceding section specific questions regarding these parameters can be satisfactorily answered. This approach is particularly useful when the values of the slope and the intercept can be correlated with chemical aspects of the problem such as the need for, or adequacy of, a blank correction or the appropriateness of a particular indicator.

From the viewpoint of routine testing, one may be interested in the calibration line as such without a critical study of specific values for the slope and the intercept. This question of practical interest can be stated as follows: How reliable is the calibration line over its entire range of applicability? The answer is obtained by the method illustrated in Figure 3, which is based on a 95% joint confidence region for the slope and intercept of the data shown in Figure 1. These data, being less precise than those of Table II, are more suitable for graphical illustration of the concepts here discussed. The two branches of the hyperbola define the

limits within which the calibration line is known at any one of its points. The line is most accurately known in the middle region of the range in which it was studied, the uncertainty of its position increasing with increasing departure from the middle. Computational details are relegated to the section on formulas and computations.

However, it is important to note at this point that the hyperbola, in addition to providing an uncertainty band for the calibration line, also yields the answer to two further classes of problems. Just as the estimated straight line can be used for the estimation of the "true" y corresponding to a given x as well as for the estimation of the x corresponding to a given y, so the hyperbola can also be used to give the confidence intervals corresponding to these two situations. The first problem is solved by drawing a vertical line through the given x; the segment of this line situated between the two branches of the hyperbola is the desired confidence interval for the "true" y corresponding to the chosen value of x. The procedure for solving the second problem is entirely analogous, the confidence interval being the segment bounded by the two branches of the hyperbola on the horizontal line drawn through the given y value. Incidentally, it is worth noting that the uncertainty intervals for x, given y, are asymmetrical with respect to the value of x situated on the calibration line. Of course, there is no impelling reason for an uncertainty interval to be symmetrical, because the uncertainty may well be greater in one direction than in the opposite one. This is the case here, the calibration line being most precisely known in the center and becoming gradually less well known at increasing distances from the center. Therefore, the uncertainty intervals for x, given y, are shorter on the side toward the center than on the other side. As may be expected, in the case in which the slope of the line is not significantly different from zero, the confidence interval for x, given y, becomes infinitely long and, of course, meaningless.

The procedure just described can be repeated for any number of given x and/or y values, using the same calibration line with its associated hyperbola, without ever causing the joint reliability of all the confidence intervals thus obtained to drop below the chosen confidence coefficient.

RELIABILITY OF QUANTITIES DERIVED FROM CALIBRATION LINE

The two classes of problems discussed in the preceding section constitute special cases of a wider class of interval estimation problems that can be solved by means of the joint confidence ellipse of slope and intercept. As will be shown in the final section, a confidence interval can be derived for any arbitrary function, linear or nonlinear, of slope and intercept. Only the linear case appears to have been considered in its most general form in the literature (*3*). An example of the nonlinear case is found in the study of the viscosity of polymer solutions.

The following equation is sometimes used to relate viscosity and concentration for dilute solutions (*11*)

$$\eta_{sp}/c = [\eta] + k'[\eta]^2 c$$

where c is concentration, η_{sp} is specific viscosity, and $[\eta]$ is intrinsic viscosity. The constant, k', which characterizes the solute-solvent system, can be estimated as the ratio of the slope, $k'[\eta]^2$, of the straight line to the square of its intercept, $[\eta]$. The uncertainty of k' is, therefore, influenced by that of both the estimated slope and the estimated intercept.

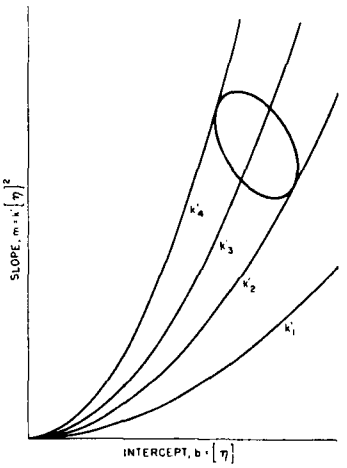


Figure 4. Confidence interval for nonlinear function of slope and intercept in viscosity study

Parabolas represent the equation $m = k'b^2$ for various values of k'. The confidence interval for k' consists of all values of k' contained between k'_2 and k'_4

The following method is proposed for solving this problem. The relation

$$k' = \frac{k'[\eta]^2}{[\eta]^2} = \frac{m}{b^2}$$

can be written $m = k'b^2$ and represents, for any given value of k', a parabola in the b, m plane (Figure 4). Different values of k' result in different parabolas some of which intersect the ellipse, while others do not. For a value of k' to be acceptable, it must correspond to a parabola which contains acceptable combinations of b and m—i.e., points inside the ellipse. Thus, the totality of acceptable values of k' is that set of k' values for which the corresponding parabolas intersect the ellipse—i.e., the values contained between k'_2 and k'_4 corresponding to the tangent parabolas. Thus, k'_1 is unacceptable, while k'_3 is acceptable.

While the problem of determining the limiting values k'_2 and k'_4 can be solved mathematically, it may be simpler in many cases, including the one under discussion, to use graphical methods involving trial and error on some values of k'.

The function of interest in the present example is the ratio of the slope to the square of the intercept. Other functions may also be of interest. If this is the case, confidence intervals can be derived for all such functions by the same general method. It can then be stated that the confidence intervals thus obtained from a single set of straight line data are all jointly valid with a probability at least equal to the selected confidence coefficient.

FORMULAS AND COMPUTATIONS

Equation of Joint Confidence Ellipse. This equation for slope and intercept can be written at once, provided the usual least squares calculations for slope and intercept are carried out in a systematic way.

Suppose that N pairs of corresponding values for x and y are given and that it is required to fit a straight line

$$y = b + mx$$

to these data. The usual assumptions are made —viz., that the x values are known without error and that the errors in the y measurements are independent of each other and have a common variance. The usual least squares formulas are then applicable and require the computation of the following quantities:

Given N pairs of x,y values, compute:

(a) From the x values:

$$S = \Sigma x \text{ and } Q = \Sigma x^2$$

(b) From the y values:

$$Y = \Sigma y \text{ and } L = \Sigma y^2$$

(c) From corresponding x and y values:

$$P = \Sigma xy$$

It is useful to represent the quantity, $NQ - S^2$, which depends on the x values only, by a separate symbol

$$\Delta = NQ - S^2 \qquad (1)$$

Then, the estimates of slope and intercept, $\hat{m}$ and $\hat{b}$, are given by

$$\hat{m} = \frac{NP - SY}{\Delta} \tag{2}$$

$$\hat{b} = \frac{QY - SP}{\Delta} \tag{3}$$

The standard error of estimate, which is a measure of the experimental error of the y measurements, is the square root of the quantity

$$s^2 = \frac{1}{N-2}\left[L - \frac{Y^2}{N} - \frac{\Delta}{N}\hat{m}^2\right] \tag{4}$$

The equation of the ellipse is

$$N(b - \hat{b})^2 + 2S(b - \hat{b})(m - \hat{m}) + Q(m - \hat{m})^2 = 2Fs^2 \tag{5}$$

In this equation, F represents the critical value of the "variance-ratio," with 2 and $N-2$ degrees of freedom, corresponding to the desired "confidence coefficient." For example, if the desired confidence is 95% and N is 10, the value of F is obtained from the "variance ratio" table at a level of significance equal to 100 − 95 = 5%, for 2 and 8 degrees of freedom. This value is 4.46.

In practice, it is not necessary actually to draw the ellipse, because it can be closely approximated by three sets of parallel tangents, as shown in the following section.

In terms of the data of Table II, the formulas just given lead to the following quantities:

$N = 10$
$S = 1943$
$Q = 676.924$
$\Delta = 2{,}991.267$
$\hat{m} = 1.00765$
$\hat{b} = 6.99$
$s^2 = 2.121$
$F = 4.46$

The equation of the ellipse is, therefore

$$10(b - 6.99)^2 + 3886(b - 6.99)(m - 1.00765) + 676{,}924\ (m - 1.00765)^2 = 18.919$$

Practical Construction of Joint Confidence Region. Referring to the symbols defined in the preceding section, the following additional quantities are required.

$$K^2 = 2Fs^2 \tag{6}$$

$$W = \frac{\sqrt{NQ}}{S} \tag{7}$$

$$L_m = K\sqrt{\frac{N}{\Delta}} \tag{8}$$

$$L_b = K\sqrt{\frac{Q}{\Delta}} \tag{9}$$

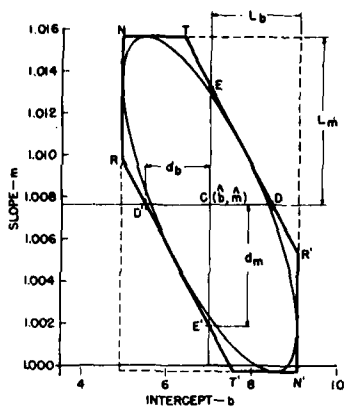


Figure 5. Graphical construction of joint confidence region

Data of Table II

$$d_m = K\sqrt{\frac{2W}{Q(1+W)}} \tag{10}$$

$$d_b = d_m\frac{L_b}{L_m} \tag{11}$$

Construct a system of coordinate axes (Figure 5) in which the abscissa represents the intercept and the ordinate the slope. The scales need not be equal for the two axes. They should be such that a rectangle of sides $2L_m$ (in the vertical direction) and $2L_b$ (in the horizontal direction) covers most of the area available for the graph. The center of the rectangle is the point, $C = (\hat{b}, \hat{m})$. After drawing the rectangle, locate the points E and E' above and below C, at distances $\pm d_m$ from C; and the points D and D' to the right and left of C, at distances $\pm d_b$ from C. Draw the sloping lines ED and $E'D'$ and extend them to their points of intersection with the rectangle. The sloping lines, as well as the four sides of the rectangle, are all tangent to the ellipse. The hexagon, $RNTR'N'T'$, enclosed between these lines inside the rectangle is an excellent approximation for the ellipse, as evidenced by Figure 5.

Formulas 6 through 11, when applied to the data of Table II, give

$K^2 = 18.919$
$W = 1.3391$
$L_m = 7.953 \times 10^{-3}$
$L_b = 2.0693$
$d_m = 5.655 \times 10^{-3}$
$d_b = 1.47$

These values were used in constructing Figure 5.

Confidence Band for Calibration Line. The hyperbola shown in Figure 3 is obtained by adding to and subtracting from the fitted value, y, corresponding to any given x, a quantity depending on the distance of this x from the average, $\bar{x}$. This quantity is

$$K\sqrt{\frac{1}{N}\left[1 + \frac{(x - \bar{x})^2}{\Delta/N^2}\right]}$$

Thus, the equation of the upper branch of the hyperbola is

$$y = \hat{b} + \hat{m}x + K\sqrt{\frac{1}{N}\left[1 + \frac{(x - \bar{x})^2}{\Delta/N^2}\right]} \tag{12a}$$

and the equation of the lower branch is

$$y = \hat{b} + \hat{m}x - K\sqrt{\frac{1}{N}\left[1 + \frac{(x - \bar{x})^2}{\Delta/N^2}\right]} \tag{12b}$$

The derivation of these formulas is outlined in the final section.

It is helpful to draw the asymptotes to the hyperbola. The equations for the two asymptotes are

$$y = \hat{b} + \hat{m}x + K\sqrt{\frac{N}{\Delta}}(x - \bar{x})$$

and

$$y = \hat{b} + \hat{m}x - K\sqrt{\frac{N}{\Delta}}(x - \bar{x}) \tag{13}$$

The quantity, $K\sqrt{\frac{N}{\Delta}} = L_m$, has already been calculated for the construction of the ellipse.

If a confidence interval is desired for y corresponding to any given value of x, say x_0, it may be determined by drawing a vertical line through $x = x_0$. The desired interval is the portion of that line which falls between the two branches of the hyperbola. Conversely, if a confidence interval is desired for x corresponding to any given value of y, say y_0, it is determined by drawing a horizontal line through $y = y_0$. The desired interval is the portion of that line which falls between the two branches of the hyperbola. It has already been pointed out that this procedure can be repeated for any number of x and y values, with the assurance that the probability that all intervals will be jointly valid is never less than the chosen confidence coefficient. In many cases, the scale of the graph will make it necessary to obtain these intervals by computation rather than graphically.

To judge the reliability of the calibration line over its entire range of applicability, it is useful to note that, for the case of equally spaced values of x, the length of the uncertainty interval for y at both extremes of the calibration line (extreme values for which measurements

were made) is approximately twice its length at an x value near the center.

Application to the data of Table II gives the following equation for the confidence hyperbola

$$y = 6.99 + 1.00765x \pm 4.35\sqrt{\frac{1}{10}\left[1 + \frac{(x - 194.3)^2}{29{,}913}\right]} \quad (14)$$

To show how this equation yields information about the precision of the calibration line

$$y = 6.99 + 1.00765x$$

let us calculate the uncertainty of a value x "read" from the line for a value of $y = 250$ mg. Substituting this value in the equation and squaring, we obtain

$$(250 - 6.99 - 1.00765x)^2 = 1.8919\left[1 + \frac{(x - 194.3)^2}{29{,}913}\right]$$

This quadratic equation in x has the roots $x_1 = 239.83$ and $x_2 = 242.48$.

This interval of uncertainty reflects only errors in the calibration curve and does not include errors in the measurement of y.

SOME THEORETICAL CONSIDERATIONS

While the basic theory of joint confidence regions has been known for a number of years, its practical usefulness seems to have been largely overlooked. Textbooks on applied statistics either ignore the issue entirely or treat it very sketchily. The present discussion is an attempt to fill this gap and to present a concise outline of the theoretical ideas and principles necessary to an understanding of the techniques already discussed in this paper. It is hoped that this exposition will also throw some light on the manner in which the various principles are related to each other.

1. It is interesting to contrast problems involving two unknown parameters, such as intercept and slope, with problems involving a single parameter. In the latter case, there is essentially only one possible confidence statement; for example, if the parameter is the slope, m, of a straight line, $y = mx$, passing through the origin, then the confidence interval for y corresponding to $x = x_0$ is the range of values extending from x_0m_1 to x_0m_2, where m_1 and m_2 are the confidence limits for m. If another value x_0' had been considered, the corresponding confidence interval would be proportional to the first, with a proportionality factor, x_0'/x_0. Thus, all such intervals are uniquely determined by values m_1 and m_2 or by each other. The joint confidence for any number of intervals thus obtained is therefore equal to the confidence for each single one, since any one interval determines all others.

On the other hand, in the case of two or more parameters, such as the slope and intercept of the line, $y = b + mx$, a confidence statement for y corresponding to x_0 does not mathematically imply a confidence statement corresponding to another value x_0'. For example, if it is stated that for $x = 2$, $b + mx$ lies between 5 and 15—i.e.,

$$5 < b + 2m < 15$$

no statement can be inferred from these inequalities for $x = 3$—i.e., for $b + 3m$. Thus, if statements of uncertainty are made both for $x = 2$ and $x = 3$, their joint reliability will be less than that corresponding to each statement taken separately. If more values of x are considered, the joint reliability will further decrease. The joint confidence ellipse ensures that no matter how many individual confidence intervals are derived from it, the joint confidence of all these intervals is never less than the selected confidence coefficient, say 0.95.

2. The usefulness of the joint confidence ellipse for slope and intercept in the study of linear relationships was recognized by Working and Hotelling (*17*) as early as 1929. Some aspects of their paper appear to have been largely ignored in subsequent writings. These authors show that the totality of all straight lines whose slopes and intercepts correspond to points inside the ellipse—i.e., the admissible lines—are contained between the branches of a hyperbola, and they point out that this hyperbola is wider than the one corresponding to the sampling errors in y for any particular x. They consider the wider hyperbola as setting limits for the "sampling errors of the trend line as a whole" (*17*). However, the ellipse discussed by Working and Hotelling involves the population standard deviation of the experimental errors, σ, and therefore, does not allow for sampling errors in the estimation of the standard error of estimate, s. Neither do these authors consider the problem of determining the uncertainty of an x value "read" from the calibration line for a given y.

3. The latter problem is examined in detail by Eisenhart (*4*). This author summarizes the theory of confidence intervals, based on Student's t, for the parameters of a straight line and the uses of hyperbolic uncertainty bands for the interval estimation of x, given y, as well as for y, given x. It will be shown that some of these problems can also be treated by the method of joint confidence regions underlying Working and Hotelling's "wider" hyperbola, leading to more satisfactory solutions for some types of applications than those based on Student's t.

4. A further aspect of linear regression is that of evaluating the uncertainty of some given function of the slope and the intercept, such as in the viscosity problem described in an earlier section. For functions that are linear with respect to b and m, a general solution based on the joint confidence ellipse is given by Durand (*3*). For functions that are of the form, $L_1(b,m)/L_2(b,m)$, where both L_1 and L_2 are linear, a solution is available using a theorem by Fieller (*5*, *7*), the basic principle of which is concisely presented by Finney (*6*). However, this solution is not based on the joint confidence ellipse and suffers, therefore, from the restriction that only one conclusion can be drawn from a given set of data, using a preselected confidence coefficient.

5. The method presented in this paper and illustrated by the viscosity problem constitutes an entirely general procedure. It contains as special cases the treatment of linear functions by the method of Durand (*3*), as well as those nonlinear functions that can be covered by Fieller's theorem. Among the latter, there is the interval estimation of x, given y which, under the generalized procedure leads also to the "wider" hyperbola of Working and Hotelling (*17*).

The principle of the general method is as follows: Given a set of data for a straight line and any arbitrary function of slope m and intercept b, say $z = f(b,m)$, first determine the joint confidence ellipse as described. Next, consider any value of z, say z_0. For this value, function $z_0 = f(b,m)$ represents a curve in the b,m plane. By varying z_0, a family of such curves is obtained. A confidence interval for z is then obtained by collecting all numerical values of z for which the corresponding curves intersect the ellipse.

This procedure can be repeated for any arbitrary number of functions of b and m, using the same ellipse. Provided that the functions contain no random errors other than those affecting estimates b and m, the confidence intervals obtained will all be jointly valid with a probability not less than the selected confidence coefficient.

6. By the procedure that has just been outlined, a single mathematical operation yields the solution to both the problem of determining a confidence band for the regression line, as dealt with by Working and Hotelling, and the problem of determining confidence intervals for y, given x, and for x, given y.

Identify z with the expression, $(y - b)/m$—i.e., x—for a fixed value of y

$$z \equiv x = \frac{y - b}{m}$$

This equation represents a straight line in the b,m plane. The values of z for which this line intersects the ellipse will be contained between two values, say x_1 and x_2, for which this line is tangent to the ellipse given by Equation 5. By means of elementary analytical geometry it can be shown that x_1 and x_2 are the solutions of the following equation in x

$$(y - \hat{b} - \hat{m}x)^2 = K^2(Nx^2 - 2Sx + Q) \quad (15)$$

The interval extending from x_1 to x_2 is, of course, a confidence interval for x for the fixed value of y considered. On the other hand, Equation 15 also represents a hyperbola entirely analogous to the "wider" hyperbola obtained by Working and Hotelling, but allowing for sampling errors in the estimate, s, of σ.

If z had been identified with the expression, $b + mx$, the same identical hyperbola would have been obtained. Thus, Equation 15 also yields confidence intervals for y, given x. By replacing the quantity, $Nx^2 - 2Sx + Q$, by its equivalent, $N(x - \bar{x})^2 + \frac{\Delta}{N}$, it is easily verified that Equations 12a and 12b are merely a different way of writing Equation 15. Figure 3 is based on these equations.

7. Fieller's theorem (*5–7*) could have been used to derive confidence intervals for x, given y, as the expression $x = (y-b)/m$ is of the form, $L_1(b,m)/L_2(b,m)$, with L_1 and L_2 linear. This technique would have led to an equation entirely similar to Equation 15, with the sole difference that quantity t^2s^2 appears in the place of K^2. As $K^2 = 2Fs^2$, the relation between the solutions obtained by Fieller's theorem and the use of the ellipse is that of the use of t *vs.* $\sqrt{2F}$. This same relation applies also to the comparison between the method described by Eisenhart, based on Student's t, and that derived from the joint confidence ellipse for confidence intervals of y, given x. In fact, it can be shown that in all cases in which confidence intervals can be derived on the basis of Student's t, they bear a constant relationship to corresponding intervals based on the joint confidence ellipse for the slope and the intercept—namely, that the length of the latter intervals is longer by the ratio, $\sqrt{2F}/t$. For a confidence coefficient of 0.95, this ratio equals 1.295 for $N = 10$ and it decreases slowly to the limiting value, 1.247, as N increases indefinitely.

Against the disadvantage of the method based on the joint confidence ellipse to yield somewhat longer confidence intervals, one must weigh two important advantages. In the first place, as mentioned before, it is possible by this method to obtain confidence intervals for any function, linear or nonlinear, of b and m. Secondly, the confidence intervals derived from the ellipse are all jointly valid, regardless of their number, with a joint confidence coefficient that is never less than the one on which the ellipse is based.

8. While the general method described above permits one to derive from a single set of data an unlimited number of jointly valid confidence intervals for y, given x, and for x, given y, it fails to solve a twofold problem discussed by others (*1, 2, 4, 12, 15*) in connection with linear regression: that of predicting in what range a "future" y, to be measured at a given x, will lie; and conversely, the problem of evaluating the uncertainty of x corresponding to a "future" y measurement. In both cases, an error is involved that is not accounted for in the joint confidence ellipse of slope and intercept—namely, the error of a future y measurement. The former problem is strictly speaking, not concerned with confidence intervals, because it deals with the uncertainty of a random variable, not a population parameter. The relation of intervals of this type to classical confidence intervals and application to a chemical example are described by Weiss (*16*).

The second problem, on the other hand, is a genuine case of interval estimation, and of particular interest to the chemist. It can be solved by an extension of the method here described through introduction of a three-dimensional ellipsoid in place of the plane ellipse (*10*). The solution thus obtained does not allow for the treatment of more than a single "future" measurement, but it does permit the construction of an unlimited number of jointly valid confidence intervals involving the slope, the intercept, and the "true" y corresponding to the "future" measured y. The method can be easily extended to include any finite number of "future" measurements.

9. Durand (*3*) notes that the use of joint confidence regions for the determination of confidence intervals for linear combinations of regression coefficients is closely related to a technique recently developed by Scheffé (*14*) and extended by Roy and Bose (*13*) for examining simultaneously all combinations of a number of observed means. The general result derived by these authors consists essentially in replacing t with $N - k$ degrees of freedom by $\sqrt{kF_{k,N-k}}$, where N is the total number of measurements and k is the number of parameters. The case discussed in this paper involves two parameters: the slope and the intercept. Making $k = 2$ in the general formula, we find $\sqrt{2F_{2,N-2}}$ as the quantity to be substituted for Student's t in the construction of confidence intervals. Since we have already found this relationship in comparing the technique based on the joint confidence ellipse with that based on Fieller's theorem in a nonlinear case, it appears that the relationship is more general than is implied in Durand's statement.

LITERATURE CITED

(1) Anderson, R. L., Bancroft, T. A., "Statistical Theory in Research," McGraw-Hill, New York, 1952.
(2) Davies, O. L., "Statistical Methods in Research and Production with Special Reference to Chemical Industry," Oliver and Boyd, London, 1947.
(3) Durand, D., *J. Am. Stat. Assoc.* **49,** 130 (1954).
(4) Eisenhart, C., *Ann. Math. Stat.* **10,** 162 (1939).
(5) Fieller, E. C., *J. Roy. Stat. Soc., Supplement* **7,** 1 (1940).
(6) Finney, D. J., *Biometrics* **5,** 335 (1949).
(7) Finney, D. J., "Probit Analysis," Cambridge University Press, Cambridge, 1952.
(8) Lark, P. D., ANAL. CHEM. **26,** 1712 (1954).
(9) Linnig, F. J., Mandel, J., Peterson, J. M., *Ibid.*, **26,** 1102 (1954).
(10) Mandel, J., unpublished work.
(11) Mark, H., Tobolsky, A. V., "Physical Chemistry of High Polymeric Systems," p. 301, Interscience, New York, 1950.
(12) Mood, A. McF., "Introduction to Theory of Statistics," McGraw-Hill, New York, 1950.
(13) Roy, S. N., Bose, R. C., *Ann. Math. Stat.* **24,** 513 (1953).
(14) Scheffé, H., *Biometrika* **40,** 87 (1953).
(15) Snedecor, G. W., "Statistical Methods," The Iowa State College Press, Ames, Iowa, 1948.
(16) Weiss, L., *Ann. Math. Stat.* **26,** 142 (1955).
(17) Working, H., Hotelling, H., *J. Am. Stat. Assoc.* **165A** (Proceedings), 73 (1929).
(18) Youden, W. J., ANAL. CHEM. **19,** 946 (1947).
(19) Youden, W. J., "Statistical Methods for Chemists," Wiley, New York, 1951.

RECEIVED for review October 3, 1956. Accepted February 11, 1957.

PREPRINT INSTRUMENT SOCIETY of AMERICA
19th Annual ISA Conference and Exhibit • October 12-15, 1964, New York
Preprint Number 12.3-2-64

UNCERTAINTIES ASSOCIATED WITH PROVING RING CALIBRATION

by Thomas E. Hockersmith
Mechanical Engineer
Harry H. Ku
Mathematical Statistician
National Bureau of Standards
Washington, D. C.

ABSTRACT

A method of error analysis is presented using data obtained from dead-weight calibration of various capacity proving rings. A breakdown of the errors into components by statistical methods and their combination into a final uncertainty statement is discussed in detail. Graphical representations are used in several places to help in the exposition.

Extension of the analysis and method of handling calibration data for multiple proving ring setups is discussed in an effort to show that the same general method of analysis should be adequate.

INTRODUCTION

A proving ring is a compact and dependable force measurement device developed at the National Bureau of Standards by H. L. Whittemore and S. N. Petrenko for the original purpose of calibrating testing machines. A typical proving ring is shown in Figure 1. It consists basically of the following components; an elastic steel ring with diametrically opposed integral loading bosses, a vibrating reed, and a micrometer dial and screw assembly. The reed and micrometer screw assembly are mounted along the diameter concentric with the bosses. When a load is applied to the ring a deflection is measured by turning the micrometer screw until positive contact is made with the vibrating reed. This deflection value is read in terms of the arbitrary scale inscribed in the face of the micrometer dial. For details on the design, use, and calibration of proving rings, see Circular of the National Bureau of Standards C 454 [1]*.

In recent years a significant increase in the use of the proving ring as a secondary transfer standard, in the field of force measurement, has precipitated the need for information dealing with accuracy of the calibration process. The purpose of this paper is to discuss methods of extracting such information from the calibration data and to present the results in a useable form.

UNCERTAINTY

Calibration may be generally thought of as the process of comparing an unknown with a standard and determining the value of the unknown from the accepted value of the standard. The accuracy of the reported values are usually given in terms of bounds to inaccuracy, or limits of uncertainty.

In any calibration process there are three possibilities available in dealing with the uncertainties. These are:

1. Report only the values obtained and make no statement about their uncertainty.
2. Make some statement of the uncertainties affecting the calibration process based on personal judgement and general experience.
3. Through the use of error analysis form an objective estimate of the uncertainties affecting the reported values.

The uncertainty of a measurement process may be characterized by giving (1) the imprecision, and (2) limits to the overall systematic error. Imprecision means the degree of mutual disagreement, characteristic of independent measurements of a single quantity, yielded by repeated applications of the process under specified conditions. The accepted unit for the imprecision of a calibration process is the standard deviation, σ, which provides a measure of how close a particular calibration result in hand is likely to agree with the results that might have been (or might be) obtained by the same calibration process in this (or other) instance(s). The larger the value of σ, the more imprecise the method of measurement, and the greater the disagreement to be anticipated between strictly comparable calibrations.

The systematic error of a calibration process refers to the more or less consistent deviations of the values observed, from the standard, or from the value intended to be measured. If the direction and the magnitude of systematic error were known with sufficient accuracy, a correction could be applied to render the reported values free from bias. Usually only limits of systematic error can be given, e.g., resulting from the uncertainty in the deter-

*The numbers in brackets refer to similarly-numbered references at the end of this paper.

mination of the mass of weights in a dead-weight load calibrating machine. Limits of systematic error are generally based on knowledge and experience with similiar measurements, information available from special studies, and judgement. In calibration the sources of systematic errors are usually studied carefully, and their effect on the final results minimized or eliminated if possible.

The total uncertainty of a calibration process places limits on its probable inaccuracy. It includes both the imprecision and the systematic error. Accuracy requires precision but precision does not necessarily imply accuracy. For example, a calibration process may be highly precise and yet when applied to a standard yield values consistently greater, or consistently less, than the accepted value of the standard.

The present method of reporting proving ring calibration employed by the NBS does not give explicitly a single expression of the overall uncertainty involved, but instead, gives estimates of the imprecision and systematic error from which the total uncertainty can be derived. This practice is in keeping with the recommendations on "Expression on the Uncertainties of Final Results" in Chapter 23 of NBS Handbook 91 [2]. The estimate of imprecision of the calibration process is given by the standard errors of the tabulated load values, which measure the combined performance of the calibration process and the particular ring. Bounds for the systematic errors are given in percent error of applied load for both dead-weight loads, and loads measured by means of a multiple ring setup.

DEAD-WEIGHT CALIBRATION

A dead-weight calibration of a proving ring consists of ten nearly equally spaced loads applied in either the 10,100-lb or the 111,000-lb capacity testing machines presently in use at the National Bureau of Standards.

Three runs of ten loads are taken on each ring to make up a calibration. Before and after each load reading a no-load reading is taken and recorded. The average of the two no-load readings is subtracted from the load reading to yield a deflection value of the ring under that load. This yields a total of thirty deflection values, three values for each load point from ten-percent of capacity to capacity load. These thirty deflections are punched on computer data cards with their corresponding load values and are fed into an electronic digital computer. A second degree equation of the form

$$D = a + bL + c(L)^2$$

is fitted to the averages of the three deflection values for each load, where

D = average deflection value
L = load in pounds

and a, b, c, are coefficients. The computer program performs the task of statistically analyzing the data, fitting the data by the method of least squares, and printing out a load versus deflection table as well as the various statistical quantities included in the report. The thirty deflection values obtained during the calibration of a 100,000-lb capacity proving ring are given in Table 1. A sample of the load versus deflection table printed out as a result of the computer fit of these data is found in Table 2.

The selection of a second degree equation in terms of load was decided upon as a result of preliminary investigation, both theoretical and experimental, to determine the proper degree of the calibration curve to represent the characteristics of the proving ring as evidenced by the raw data. At the same time it was necessary to keep in mind the many problems associated with applying an error analysis to such data.

Figure 2 shows the three deflection values at each of the ten load points for proving ring A, with most of the linear trend removed from the deflection values. The smooth curve represents the plot of the computed deflection values derived from the second degree fit with the same linear trend removed. This figure shows how well the second degree curve fits the observed deflections.

Several interesting and useful comparisons resulting from the error analysis and fitting techniques employed are as follows. From the dispersion of the three deflection values at each load point about their average, the standard deviation of a deflection value can be computed with two degrees of freedom. Since these standard deviations computed over the range of loads are comparable in magnitude, the ten values may be pooled together. This pooled value of the standard deviation, denoted as s_w, can be compared to its long run average value over many previous calibrations to determine if the calibration process is under control, i.e., stable with respect to precision.

A standard deviation s associated with the calibration of this particular ring can be computed from the residuals of the ten average deflection values about the second degree curve. This value of the standard deviation, s, can be compared with the pooled standard deviation of an average deflection value, $s_w\sqrt{3}$, obtained from the ten sets of triplicate deflection values (i.e., the pooled estimate of the standard deviation of an individual deflection divided by $\sqrt{3}$). If the two standard deviations s and $s_w\sqrt{3}$ are of

nearly the same magnitude then the ring is in good condition and the scatter of the points is due mainly to the inability of the calibration process as a whole to repeat. Conversely, if the standard deviation s computed from the deviations of the ten average deflection values from the curve is considerably larger than $s_{\bar{y}}\sqrt{3}$, the estimated standard deviation of an average deflection value, then the condition of the ring is not good and reconditioning by the manufacturer is indicated. An example of this can be seen in figure 2. This ring is apparently not in good condition since the broken curve connecting the averages of the three deflection values at each load point does not follow the fitted curve closely. For rings in good condition, the two curves are practically indistinguishable on a graph to this scale. In the future this type of reasoning may be used as a basis for acceptance or rejection of a particular device.

Previously a calibration graph was included with the calibration certificate as shown in figure 3. This graph was a plot of the calibration factor for the ring in pounds per division versus the deflection in divisions. The straight line through the points was drawn for "best fit". Because the calibration factors were computed by dividing each deflection into its corresponding load, the points of the plot near the lower end of the load range, of the device, show considerably more dispersion than the points near the upper end. Therefore the upper points were considered to be better indicators for the drawing of the "best fit" line through the plotted points. In the case of the second degree fit of deflection versus load by the method of least squares, the individual points are treated with equal weight, a more accurate fit of the calibration data is obtained, and no possibility of personal bias is introduced.

The above can be illustrated as the by-products of a simple test designed and suggested by W. J. Youden of NBS. This test consisted of several operators taking readings with a proving ring under various known dead-weight loads. These loads were then computed as if they were unknown using first the table of load values from the second degree fit and second the load values derived from the "best fit" curve. Comparison showed that over the range of the ring, the load values computed from the second degree fit were closer to the actual known loads applied to the ring. Therefore, if the ring is to be used over its entire calibrated range the second degree fit gives more accurate load values. The same data were also used to check the computed limits of uncertainty for the particular ring and in no case did the difference between the actual and computed load value exceed these limits. A sample of the values determined during this experiment can be found in Table 3.

In order to arrive at some measure of dependability of the values given in the load table the corresponding confidence interval is needed. To determine such an interval, the standard error of a deflection value for a given load is computed and some multiple of this value is used as limits of uncertainty on the imprecision.

To predict a deflection value D_i for a particular given load L_i, the deflection value can be expressed as $D_i = a + bL_i + cL_i^2$. Thus D_i is a linear combination of the coefficients estimated, and its standard error s_i can be expressed in terms of the standard deviation s (estimated from the residuals of the fit, with seven degrees of freedom) and the load L_i, and the variance-covariance matrix $[C_{ij}]$ of the estimated coefficients a, b, and c, as follows:

$$s_i^2 = \underline{L}'[C_{ij}]\,\underline{L}\,s^2$$

where the vector $\underline{L}' = (1, L_i, L_i^2)$.
(For details of the method of polynomial fitting used and the calculation of the standard error s_i the reader is referred to sections 6-3 and 6-5 of Chapter 6 of NBS Handbook 91 "Experimental Statistics" [2].)

The dependence of s_i on the value of L_i indicates that D_i values corresponding to L_i values at the two ends of the range of L have larger prediction errors than do D_i values corresponding to L_i in the center portion. For convenience, the largest value of the standard error s_i computed from the above expression is used for all values of L_i in a proving ring report, and for ten equal increments of equally spaced loads L_i, the value of the largest standard error, s_i, is approximately equal to 0.79s. This is converted into load in pounds by multiplying 0.79s by the *maximum calibration factor, in pounds per division*, for the particular ring.

Using the t statistic and the computed standard error a confidence interval for the deflection value on the curve for a single given load can be calculated. In calibration work, however, we require not merely the calculation of a confidence interval for the deflection value corresponding to a single load, but the calculation of a confidence band for the whole calibration curve. Therefore, a wider interval will be required for the same level of confidence. The confidence band for a line as a whole is discussed on pages 5-15 to 5-17 of reference [2] and for entire curves, in Chapter 28 of [3], where it is shown that in the general case, the half width at $L = L_i$ of the confidence band for the curve as a whole is:

$$\sqrt{k\,F_{.05}(k,\nu)} \times s_i ,$$

where k is the number of coefficients estimated, ν is the number of degrees of freedom in estimating s_i, and F is an appropriate upper percentage point of the distribution of the F statistic (as an illustration we are using the upper 5%

point). Thus for $k = 3$, $\nu = 7$, and ten equal increments of loads the half width of the 95% confidence interval is $\sqrt{3 \times 4.35} \times 0.79s = 2.86s$. (Since the largest value of the computed standard error is used, the confidence level is at least 95%.) Therefore the over-all limits of uncertainty for the calibration by this procedure could be expressed as 2.86 x s, (s is the standard deviation given in the report), plus the systematic error.

It may appear that the above procedure for determining the limits of uncertainty in the calibration of a proving ring by basing it on the prediction of a deflection value for a given load is a reverse procedure. However, for the method of calibration described this seemingly reverse procedure is the proper one. Figure 4 is a schematic diagram of the deflection - load curve obtained from a calibration with the corresponding confidence band sketched about it. For any given load, the true deflection value is expected to be situated within the band. Conversely, if a deflection value d is given, a horizontal line parallel to the load axis will intercept the curve at the corresponding load value $\hat{L}$; in addition this line will also intercept the band at two points L_i and L_u which give the corresponding lower and upper confidence limits for the load. This is true provided that the deflection value is known without error. If the uncertainty of the deflection value can be represented by D_i and D_u, then the corresponding confidence interval for the load will be wider, as given by L_i' and L_u'. In other words, the accuracy with which the deflection readings are obtained in using the ring must be taken into account by the user of the ring.

Each load value given in the table of load versus deflection is therefore the predicted value of the load, given a deflection value, and is expected to be within the uncertainty limits given for the calibration.

CALIBRATION OF RINGS USING MULTIPLE RING SETUPS

The present practice for calibrating proving rings with nominal capacities in excess of 110,000 lb is as follows:

1. to divide the nominal capacity into ten approximately equal increments,
2. to calibrate the ring by dead weights for the increments of load less than 110,000 lb, and
3. to calibrate the ring by either a 3, 4, or 5 proving ring setup for the remaining increments of load.

For a calibration using this procedure there are a number of problems relating to the analysis of data and interpretation of results. Some of these problems cannot be solved without considerable changes in the procedure of calibration. Since such changes are impractical, and in the near future dead-weight calibration capacity will be extended to 1,000,000 lb, one solution is to fit multiple ring calibrations by the same method as that for the dead-weight calibrations. The following discussion is based on the results of calibrations of rings fitted by this method.

Examination of these results showed no evidence of bias in the sense that residuals of the fit at the two adjoining increments of load, i.e. the last dead-weight and the first multiple ring load, are not unusually large or consistently of opposite sign. For this to remain true it is necessary that the calibrations of the rings used to determine the load in a multiple ring setup be unbiased. To insure that this condition is maintained the rings owned by the Bureau are usually reconditioned yearly and are calibrated frequently.

For dead-weight calibration the errors of the applied loads were assumed to be negligible in fitting the data; for multiple ring calibration, errors are introduced in the determination of the loads applied. Thus a non-linear functional relationship is to be estimated between deflection and load where the measurements of both are subject to error. There is no simple solution to this problem except that experience in this laboratory has shown that the least square fitting procedure still gives satisfactory estimates provided the errors are small compared with the range covered. This requirement is satisfied since each increment of load is more than 700 times the magnitude of the error involved.

Considering the above, and from a study of numerous past calibrations, it was decided the deflection should be fitted as a function of load since the former is believed to have larger errors than the latter.

Since the dead-weight calibration is presumably more precise than the multiple ring calibration, the question of weighting the observations prior to least square fitting was considered. Results of the rings studied indicated that the standard deviation of an average deflection obtained from multiple ring calibrations was not significantly larger than that for the dead weights. Thus the inflation of this deviation due to the errors in the loads does not increase the total imprecision by any appreciable amount. The use of weighting factors is therefore not of practical importance.

Examination of the plot of residuals resulting from fitting deflections to the loads, both in dead weights and in multiple ring setups, indicates that the deviations of the data points

from the fitted curve contribute a large part of the total error. In view of this it appears reasonable that the averages of the deflection readings should be used for fitting, similar to the procedure used for dead-weight calibration. Thus, the standard error includes the imprecision components of the calibration error for both the ring being calibrated and the rings being used to measure the applied load.

Bounds for systematic error of a multiple ring setup can be estimated by summing (1) the systematic error due to the dead weights (2) the systematic difference due to change with time in the calibrated values of the load measuring devices, and (3) other sources of error due to the inherent difficulties in using and reading the devices simultaneously. For example, such an estimate can be given an percent error of applied load for loads in excess of the dead weights.

CONCLUSION

In the above we have presented a procedure for the determination of limits of uncertainty for the calibration of proving rings. The method of analysis includes; the fitting of this type of data to an appropriate curve by the method of least squares, the use of confidence intervals and bands as limits of imprecision, and the estimate of bounds for systematic error.

Since many types of devices and instruments are calibrated similiarly at selected points along their ranges, it is believed that the procedures outlined above may be useful, when properly modified, in yielding a realistic evaluation of the uncertainties associated with their calibration.

ACKNOWLEDGMENT

Acknowledgment is made to Churchill Eisenhart of the National Bureau of Standards for the many helpful discussions and suggestions.

REFERENCES

(1) B. L. Wilson, D. R. Tate, and G. Borkowski, "Proving Rings for Calibrating Testing Machines", NBS Circular C 454, U. S. Government Printing Office, Washington 25, D. C.

(2) M. G. Natrella, "Experimental Statistics," NBS Handbook 91, U. S. Government Printing Office, Washington 25, D. C.

(3) M. G. Kendall and Alan Stuart, Advanced Theories of Statistics, Hafner Publishing Co., New York, 1961.

Figure 1 Proving Ring

Table 1 A Calibration of Proving Ring A

Applied load lb	Deflection Run 1 div	Run 2 div	Run 3 div
10,000	68.32	68.35	68.30
20,000	136.78	136.68	136.80
30,000	204.98	205.02	204.98
40,000	273.85	273.85	273.80
50,000	342.70	342.63	342.63
60,000	411.30	411.35	411.28
70,000	480.65	480.60	480.63
80,000	549.85	549.85	549.83
90,000	619.00	619.02	619.10
100,000	688.70	688.62	688.58

Table 2 - Computed Load Table in lb for 70 Degrees F for Proving Ring A

Deflection Div	0	1	2	3	4	5	6	7	8	9
60.	-	-	-	-	-	-	-	-	9952.	10099.
70.	10245.	10392.	10538.	10685.	10831.	10978.	11124.	11270.	11417.	11564.
80.	11710.	11856.	12003.	12149.	12295.	12442.	12588.	12735.	12881.	13027.
90.	13174.	13320.	13467.	13613.	13759.	13906.	14052.	14199.	14345.	14491.
100.	14638.	14784.	14930.	15077.	15223.	15369.	15516.	15662.	15808.	15954.
-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-
640.	93007.	93151.	93295.	93439.	93582.	93727.	93871.	94014.	94158.	94302.
650.	94446.	94590.	94734.	94878.	95021.	95165.	95309.	95453.	95597.	95741.
660.	95885.	96029.	96173.	96316.	96460.	96604.	96748.	96892.	97035.	97179.
670.	97323.	97467.	97611.	97754.	97898.	98042.	98186.	98330.	98473.	98617.
680.	98761.	98905.	99048.	99192.	99336.	99480.	99623.	99767.	99911.	100054.

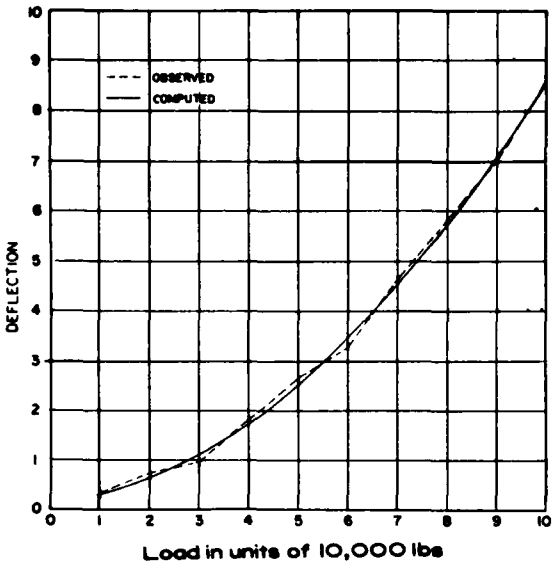


Figure 2 Observed and Fitted Calibration Curves for Proving Ring A

Note: Deflection = Deflection value minus 68.00 x (the number of the load increment)

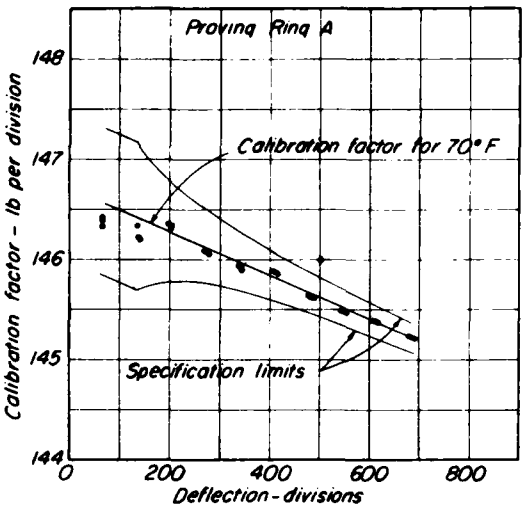


Figure 3 Calibration Graph for Proving Ring A

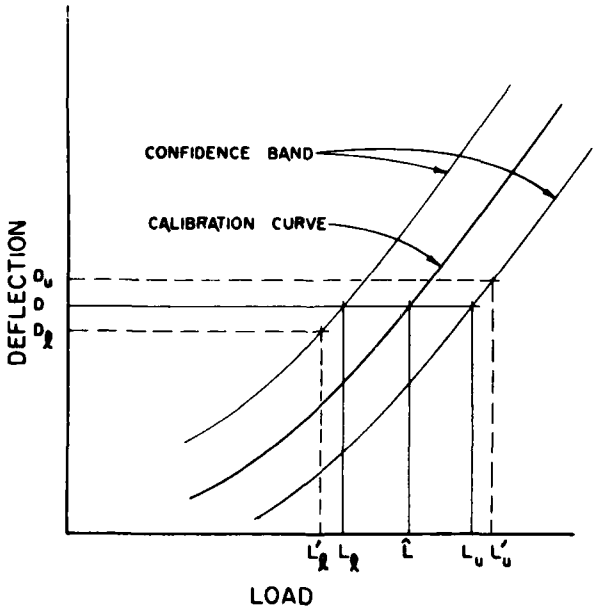


Figure 4 Determination of Confidence Limits for a Load given a Deflection Value

Table 3 - Sample Results of Experiment Designed by W. J. Youden of NBS for Proving Ring A

A Load applied to ring lb	B Computed load using second degree fitting method lb	Column A minus Column B lb	C Computed load using "best fit" method lb	Column A minus Column C lb	Ring reader
10,070	10,077	- 7	10,090	-20	1
30,000	29,987	+13	29,994	+ 6	1
40,050	40,059	- 9	40,064	-14	1
80,020	80,032	-12	80,036	-16	1
10,020	10,033	-13	10,046	-26	2
30,050	30,047	+ 3	30,053	- 3	2
40,000	40,011	-11	40,016	-16	2
80,070	80,081	-11	80,082	-12	2
10,000	9,993	+ 7	10,007	- 7	3
30,070	30,063	+ 7	30,069	+ 1	3
40,020	40,029	- 9	40,034	-14	3
80,050	80,061	-11	80,062	-12	3
10,050	10,058	- 8	10,068	-18	4
30,020	30,010	+10	30,013	+ 7	4
40,070	40,087	-17	40,096	-26	4
80,000	80,025	-25	80,030	-30	4

NOTE: An edited version of this paper has been published in the ISA Journal, Instrument Society of America, Vol. 12, No. 6, June, 1965.

Reprinted from Journal of the Washington Academy of Sciences 54, 24-33 (1964).

The Meaning of "Least" In Least Squares*

Churchill Eisenhart

National Bureau of Standards

I. Introduction

The present status of the Method of Least Squares is this: Everyone uses it, but not in exactly the same way, nor for the same reasons. There is thus some similarity to the present status of Probability, with respect to which Bertrand Russell has remarked (1): "While interpretation in this field is controversial, the mathematical calculus itself commands the same measure of agreement as any other branch of mathematics." But the situation with respect to the Method of Least Squares is not exactly parallel: In the case of the Method of Least Squares there is complete agreement on the procedure for forming the 'normal equations' from the fundamental 'observational equations,' and everyone comes up with the very same numbers for the solutions of the normal equations; but their reasons for employing the Method of Least Squares, their understanding of its objectives and the conditions under which these are achieved, and their interpretations of end results of its application, may be quite different. Furthermore, in contrast to the situation in Probability, individuals who utilize the 'Method of Least Squares' as a tool in their own line of work are usually not aware of the existence of alternative formulations of this technique.

This somewhat extraordinary situation results from the fact that the Method of Least Squares was developed originally from three distinctly different points of view: (1) *Least Sum of Squared Residuals* (Legendre, 1805), (2) *Maximum Probability of Zero Error of Estimation* (Gauss, 1809), and (3) *Least Mean Squared Error of Estimation* (Gauss, 1821). These differ not only in their aims and in their initial assumptions, but also in the meanings that they attach to the numbers that all three yield as a common answer to any given problem. Unfortunately, the existence of these three different formulations and consequent different interpretations of the end results of applying 'Least Squares' are rarely mentioned in books on the practical application of the Method of Least Squares. The only exception in English of which I am aware is Whittaker and Robinson's *The Calculus of Observations* (2), first published in 1924: chapter IX contains a discussion of Legendre's original formulation, in which no probability considerations are involved; a full treatment of Gauss's first "proof," in which what we now term the 'normal distribution' plays a central and indispensable role; and a brief summary of Gauss's second development, which he showed to be independent of the functional form of the law of error involved whenever the 'best values' implied by the techniques of *Least Sum of Squared Residuals* are linear functions of the basic observations. Gauss himself decidedly preferred his second formulation, the existence of which seems to be virtually unknown to almost all American users of "Least Squares," except students of advanced mathematical statistics.

* Extracts from a paper in preparation on "The Background and Evolution of the Method of Least Squares."

II. Minimization of Residuals and Legendre's "Methode des Moindres Quarres"

The Method of Least Squares evolved early in the 19th century in response to a recognized need for a 'best' general procedure for the combination of observations in astronomy and geodetic surveying.

When two or more related quantities are measured individually, the resulting measured values usually fail to satisfy the constraints on their magnitudes implied by the given interrelations among the quantities concerned. In such cases these 'raw' measured values are mutually contradictory and require 'adjustment' in order to be usable for the purpose intended.

Inasmuch as the actual *errors* of individual observations are usually unknown and forever unknowable, the early attempts to achieve a good adjustment seem to have concentrated on minimizing the apparent inconsistency of a set of observations as evidenced by some simple function of their *residuals*.* The practical requirements of unique solutions and computational simplicity then led, in due course, to the technique of *Least Sum of Squared Residuals*. This was the essence of Legendre's "Méthode des Moindres Quarrés," proclaimed in 1805 (3). No probability considerations were involved.

The successive stages of this evolution of the Method of Least Squares were:

1. When several 'equally good' measurements of a single quantity were available, the Principle of the Arithmetic Mean stated that the 'best' value to take was their arithmetic mean. The arithmetic mean a of a set of measurements $Y_1, Y_2, \ldots, Y_n$ is the solution of the equation

$$\sum_{1}^{n}(Y_i - a) = 0, \qquad (1)$$

that is, the value determined by the *condition of zero sum of residuals*.

This principle seems to have originated in western Europe sometime in the latter half of the 16th century A.D. and appears to have evolved from the technique of taking measurements in pairs such that the two members of a pair are affected by systematic errors of (approximately) equal magnitude but of opposite signs, in which case the arithmetic mean of a pair is (at least, more nearly) free from the effects of these errors.

2. Roger Cotes (1682-1716), in his *Aestimatio errorum* (4), suggested that, when several determinations of a single quantity were available that were subject to unequal uncertainties, then the 'best' value to take for the quantity in question is the weighted arithmetic mean of the individual determinations weighted "inversely proportional to the lengths of the Deviations over which one can spread [their] Errors."

3. Application of Cotes's suggestion to determining the slope β of a line through the origin, $y = \beta x$, from observational points $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$ affected by errors in the y-direction only, leads to taking the value B determined by the equation

$$\sum_{i=1}^{n}(Y_i - Bx_i) = O \qquad (2)$$

as the 'best' value for β, when the uncertainties of the respective Y_i are essentially constant over the range of value of x involved. If the Y_i are regarded as observed values of the respective quantities βx_i, for which the corresponding adjusted values are Bx_i, $(i = 1, 2, \ldots, n)$, then (2) clearly expresses *the condition of zero sum of residuals;* and, when written in the form

$$\bar{Y} - B\bar{x} = 0, \qquad (3)$$

* If $Y_1, Y_2, \ldots, Y_n$ are observed values of a magnitude α, then $Y_1 - \alpha = E_1$, $Y_2 - \alpha = E_2, \ldots, Y_n - \alpha = E_n$ are the *errors* of the respective observations. If, the value of α being unknown, one adopts some particular value for it, say a, then $Y_1 - a = R_1$, $Y_2 - a = R_2$, $\ldots$, $Y_n - a = R_n$ are the *residuals* of the observations corresponding to the *adjusted value* a.

where Y and x are the arithmetic means of the Y- and x-values respectively, shows that "the Cotes line," $y = Bx = (\overline{Y}/\overline{x})x$, passes through the two-dimensional center-of-gravity of the data, $(\overline{x}, \overline{Y})$.

4. In 1748, Leonard Euler (1707-1783) and Tobias Mayer (1723-1762) independently devised and applied (5, 6) an extension of the condition of zero sum of residuals to multi-parameter problems that is today called the Method of Averages: this consists of subdividing the observational points into as many subsets as there are coefficients to be determined, the subdivision being in terms of the values of (one of) the independent variable(s), and then applying the condition of zero sum of residuals to the points of each subset, in the manner of equation (2) above. Provided that one is thus able to form as many distinct observational subsets as there are unknown parameters to be determined, the Method of Averages will always come up with a value for each parameter. But there is usually some arbitrariness and room for subjective choice in the formation of the subsets, with consequent variation in the answers obtained.

5. As a means of overcoming such arbitrariness and subjectivity, Roger Joseph Boscovich (1711-1787) proposed that, given more than two pairs of observed values of variables x and y connected by a linear functional relationship of the form $y = \alpha + \beta x$, then the values (a and b) that one should adopt for α and β in order to obtain the line ($y = a + bx$) that is most nearly in accord with all of the observations should be those determined jointly by the two conditions:—

I. *The sums of the positive and negative residuals (in the y-direction) shall be equal.*

II. *The sum of the absolute values of all of the residuals shall be as small as possible.*

Condition I implies that the best fitting line $y = a + bx$ shall pass through the centroid $(\overline{x}, \overline{y})$ of the observational points. Condition II *in conjunction with Condition I* requires that the slope b shall satisfy the equation

$$\sum_{i=1}^{n} | (y_i - \overline{y}) - b(x_i - \overline{x}) | = \text{minimum.} \quad (4)$$

Consequently, determination of a "Boscovich line" reduces to determining its slope b from equation (3) and then evaluating a from the relation $a = \overline{y} - b\overline{x}$.

Boscovich stated and applied his two conditions for a line of best fit for the first time in his 1757 summary and reevaluation (7) of the measurement of a meridian arc near Rome by Christopher Maire and himself, first published in 1755. In this first pronouncement and application of his method he does not give any indication of how he solved equation (4) to obtain the 'best' value of the slope b. Three years later (8), Boscovich restated his two conditions and then gave a very useful algorithm for solving equation (4), together with a geometric proof of its validity, followed by a step-by-step illustration of its application. His algorithm and his proof, in outline, may be found in my chapter in the Boscovich Memorial Volume edited by L. L. Whyte (9).

6. Pierre Simon, Marquis de Laplace (1749-1827), in his first memoir on the Figure of the Earth (10), proposed, as a test of the adequacy of a linear relation $y = a + bx$ to describe a given set of data, that the values of a and b be chosen so as to *minimize the absolute value of the largest deviation* and then a subjective judgment made whether the resulting largest residual is, or is not, explainable in terms of the recognized uncertainties of the data involved. He also outlined a procedure for determining the required values of a and b. In his second memoir on the Figure of the Earth (11), Laplace adopted Boscovich's two criteria for a line of best fit and gave an algebraic formulation and derivation of Boscovich's algorithm for solving equation (4) above. In Book III, Chapter 5, of his Mécanique Celeste (12),

Laplace described again (pp. 417-424) the method that he had used in 1783 to determine the line that minimizes the absolute value of the maximum residual and then gave (pp. 424-434) an alternative procedure for achieving the same end "when the number of observations is considerable." He also extended (pp. 438-442) his 1789 algebraic formulation of Boscovich's technique to the case of observational points of unequal weight.

7. In 1795, at the age of eighteen, Carl Friedrich Gauss (1777-1855), mathematical peer of Archimedes (287-212 B.C.) and Sir Isaac Newton (1642-1727) and unequaled in mathematical precocity, discovered the algebraic and arithmetical advantages of the technique of *Least Sum of Squared Residuals* for adjustment of observations in geodesy.

"Originally Gauss did not attach great importance to the method of least squares; he felt it was so natural that it must have been used by many who were engaged in numerical calculations. Frequently he said that he would be willing to bet that elder Tobias Mayer (1723-1762) had used it in his calculations. Later he discovered by examining Mayer's papers that he would have lost the bet." (13. p.113).

This may serve to explain in part why Gauss did not publish anything on the Method of Least Squares for over a decade, although he employed the Method almost daily from 1801 onwards in a great variety of astronomical calculations. (14, p. 98).

8. Adrien Marie Legendre (1752-1833) introduced the world to the technique of *Least Sum of Squared Residuals* in his book on "New Methods for Determining the Orbits of Comets" (3) published in 1805. In an Appendix "On the Method of Least Squares," occupying pages 72-80, he wrote:

"Of all the principles which can be proposed for [the combination of observations] I think there is none more general, more exact, and more easy of application, than that of which we have made use in the preceding researches, and which consists of rendering the sum of the squares of the errors as a *minimum*. By this means there is established among the errors a sort of equilibrium which, preventing the extremes from exerting an undue influence, is very well fitted to reveal that state of the system which most nearly approaches the truth."

Legendre then proceeded to deduce his now well-known rules for forming the so-called 'normal equations.' He then shows that the Principle of the Arithmetic Mean is a special case of his Principle of *Least Sum of Squared Residuals.*

Unfortunately, throughout Legendre's exposition of his "Méthode des moindres quarrés," and his illustrations of its application, he used the term "errors" for what are more accurately termed *residuals.* This has served to confuse the unwary and to conceal the distinction between what he merely asserted in 1805 and what Gauss showed in 1821 to be a statistical property of the procedure. The essence of what Legendre said is this: If in the interest of achieving an objective adjustment one seeks to minimize the mutual inconsistencies of the observations as measured by some simple function of their *residuals,* then the practical requirements of general applicability, unique arithmetical solutions, and ease of computation lead to the adoption of the technique of *Least Sum of Squared Residuals.* No probability considerations were involved. And his "discovery" simply marked the culmination of the attempts by Euler, Mayer, Boscovich, Laplace, and others to develop a practicable objective method of adjustment based solely on consideration of residuals.

III. 'Laws of Error' and Gauss's First 'Proof' of the Method of Least Squares

The *error* of a measurement Y is, by definition, the difference $Y - \tau$ between the measurement and the *true value* τ of the quantity measured. The error of a particular measurement, y, is, therefore, a fixed number, $y - \tau$. The numerical magnitude and sign of this number are ordinarily unknown and unknowable, because τ, the true value of the quantity concerned, is usually unknown and un-

knowable. A mathematical *theory of errors* is not possible so long as individual measurements are regarded as unique entities, that is, as *fixed numbers* $y_1, y_2, \ldots$. A mathematical theory of errors is possible only when particular measurements $y_1, y_2, \ldots$ are regarded as instances characteristic of the measurements $Y_1, Y_2, \ldots$ that might have been, or might be, yielded by the same measurement process under the same circumstances. This fundamental step was taken on March 4, 1755, by Thomas Simpson (1710-1761), Professor of Mathematics at the Woolwich Military Academy, in "A Letter to the Right Honourable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations, in practical astronomy" (15). This remarkable letter began as follows:

"My lord, it is well known to your lordship, that the method practiced by astronomers, in order to diminish the errors arising from the imperfections of instruments, and of the organs of sense, by taking the Mean of several observations, has not been so generally received, but that some persons, of considerable note, have been of opinion, and even publicly maintained, that one single observation, taken with due care, was as much to be relied on as the Mean of a great number.

"As this appeared to be a matter of much importance, I had a strong inclination to try whether, by the application of mathematical principles, it might not receive some new light; from whence the utility and advantage of the method in practice might appear with a greater degree of evidence. In the prosecution of this design (the result of which I have now the honour to transmit to your Lordship) I have, indeed, been obliged to make use of an hypothesis, or to assume a series of numbers, to express the respective chances for the different errors to which any single observation is subject . . .

"Should not the assumption, which I have made use of, appear to your Lordship so well chosen as some others might be, it will, however, be sufficient to answer the intended purpose: and your Lordship will find, on calculation that, whatever series is assumed for the chances of the happening of the different errors, the result will turn out greatly in favour of the method now practised, by taking a mean value."

Simpson's first "hypothesis" was that the *errors* of measurements of a single quantity by a particular measurement process be regarded as taking the values $-v, -v+1, \ldots, 2, 1, 0, 1, 2, \ldots, v-1, v$, with equal probabilities, *i.e.*, a *discrete uniform distribution.* Next, he assumed that the errors be regarded as taking on the above values with probabilities proportional to $1, 2, \ldots, v-1, v, v+1, v, \ldots, 2, 1$, respectively, *i.e.*, a *discrete isosceles triangle distribution.* Utilizing the generating function techniques that had been employed by Abraham DeMoivre (1667-1754) for the solution of problems relating to tosses of dice and other games of chance (16), Simpson derived, for each of these distributions, the probability distribution of the *sum* of *n independent errors* from such a distribution, and then from these the corresponding distributions of the *arithmetic mean* of *n independent errors.* He summed up his findings as follows:

"Upon the whole of which it appears, that the taking of the Mean of a number of observations, greatly diminishes the chances for all the smaller errors, and cuts off almost all possibility of any great ones: which last consideration, alone, seems sufficient to recommend the use of the method, not only to astronomers, but to all others concerned in making of experiments of any kind (to which the above reasoning is equally applicable). And the more observations or experiments there are made, the less will the conclusion be liable to err, provided they admit of being repeated under the same circumstances."

In a second paper on "the advantage arising by taking the mean" (17), Simpson found the distribution of the mean of *n* independent errors from a *continuous isosceles triangle distribution,* by proceeding to the limit as the spacing between the error values in the fixed interval $(-a, +a)$ tends to zero.

It should be noted that Simpson did *not* prove that "taking of the arithmetic mean" was the best thing to do, but merely that it *is* advantageous. However, in accomplishing this goal he did something

much more important: he took the bold step of regarding errors, not as individual unrelated happenings, but as properties of the measurement process itself and the observer involved. He thus opened the way to a mathematical theory of measurement based on the mathematical theory of probability.

Simpson's idea of probability distributions of error was taken up quickly on the Continent. Joseph Louis, Comte de Lagrange (1736-1813), an Italian by birth, a German by adoption, a Frenchman by choice, and one of the greatest mathematicians of all time, reproduced and elaborated on Simpson's results—without mention of Simpson—in a long memoir "on the utility of taking the mean" (18).

Without a similar passage to the limit he deduced the (subsequently oft rediscovered) distribution of the arithmetic mean of n independent errors from a *continuous uniform distribution.*

Daniel Bernoulli (1700-1782), nephew of James Bernoulli (1654-1705) whose *Ars Conjectandi* (1713) is one of the great landmarks in the history of probability, published in 1778 a highly original paper on "The most probable choice between several discrepant observations and the formation therefrom of the most likely induction" (19) that apparently existed in manuscript as early as 1774 (20, p. 634). In this paper Bernoulli proposed (1) a *semi-circular law of error,*

$$f(x) = \frac{2}{\pi a^2}\sqrt{a^2 - x^2}, \quad -a \lessgtr x \lessgtr +a,$$

where $x = y - \tau$ is the error of y as an observed value of the *true value* τ, and $\pm a$ are limits which an error will never exceed; and (2) advocated maximization of the product $f(x_1)f(x_2) \ldots f(x_n) = \left(\frac{2}{\pi a^2}\right)^n \prod_{i=1}^{n} [a^2 - (y - \tau)^2]^{1/2}$ with respect to τ to obtain the "most likely value" of τ indicated by the observations $y_1, y_2, \ldots, y_n$. Today we would call this "most likely value", $T = T(y_1, y_2, \ldots, y_n)$, the *maximum likelihood estimate* of τ corresponding to the law of error $f(x)$. For $n = 3$, evaluation of T requires the solution of an equation of the fifth degree consisting of twenty terms; and for $n > 3$, the algebra and arithmetic become unmanageable. However, for $y_1 \lessgtr y_2 \lessgtr y_3$, Bernoulli showed that his "most likely value" T is greater than, equal to, or less than the arithmetic mean of the three values according as the middle value (y_2) is less than, equal to, or greater than the midpoint $\frac{1}{2}(y_1 + y_3)$ between the extremes, respectively. His T thus assigns greater weight to the more distant of the two extreme observations. The actual magnitude of the difference $T - \bar{x}$ depends, however, on the choice of a, the limit an error will never exceed in absolute value, but tends to zero rapidly as $a \to \infty$, leading Bernoulli to remark: "Those who are most shocked by our principles will have nothing further to contradict if only they make the field of possible deviations as large as possible."

In 1774, Laplace, in his first discussion of the problem of the 'best mean' (20), proposed (1) *a double-exponential law of error,*

$$f(x) = \frac{m}{2} e^{-m|x|}, \quad -\infty < x < +\infty;$$

and (2) adoption as the 'best mean' that function $T(Y_1, Y_2, Y_3)$ of three observations for which the average value of $|T - \tau|$ is a minimum. Today we would call his T the *minimum mean absolute error estimator of* τ. *For* $n = 3$ *and* $y_1 \lessgtr y_2 \lessgtr y_3$, Laplace's 'best mean' T is greater than, equal to, or less than y_2, the middle value (*i.e.*, the *median*), according as y_2 is less than, equal to, or greater than $\frac{1}{2}(y_1 + y_3)$, the midpoint between the extremes, respectively. T is thus a 'corrected median', the correction being in the direction of the more distant of the two extreme observations. Furthermore, $T \to y_2$ as $m \to \infty$ (*i.e.*, very high *precision*); and $T \to y$, the mean of the three values, as $m \to 0$ (*i.e.*, very poor precision).

Thus, while Simpson's and Lagrange's work had shown the arithmetic mean to be increasingly 'good' as $n \to \infty$, Bernoulli's and Laplace's work implied that the arithmetic mean was 'best' only in the limiting case of infinitely poor precision.

As noted above, Gauss discovered the great algebraic and arithmetical advantages of the technique of *Least Sum of Squared Residuals* in 1795. In 1797 he attempted to justify this procedure via the calculus of probabilities, concluding that determination of "most probable values" of unknown quantities is impossible unless the law of error is known explicitly. "When this is not the case, nothing remains but to assume such a function as an hypothesis. It seemed to him most natural to proceed first the other way around and to look for that function on which the whole theory should be based if for the simplest case there is to result the rule generally accepted as good, namely, that the arithmetic mean of several values obtained for the same unknown through observations of equal reliability is to be considered as the most probable value" (14, p. 98). By June 1798 (13, p. 113) he had completed his now famous 'proof' of the Method of Least Squares, in which he (a) adopted as a postulate the Principle of the Arithmetic Mean, (b) utilized the concept that repetition of a measurement process generates a *probability distribution of errors*, and (c) applied Bayes's *method of inverse probability*—without reference to Thomas Bayes (1702-1761). Starting from these premises he showed that if the arithmetic mean of n independent measurements of a single magnitude is to be the most probable value of this magnitude *a posteriori*, then the *errors* $X_i = Y_i - \tau$ of the individual measurements Y_i must be distributed in accordance with the law of error

$$f(x) = \frac{h}{\sqrt{\pi}} e^{-h^2x^2} \qquad -\infty < x < +\infty.$$

(5) Then he showed that, *if* errors are normally distributed, and *if* the unknown values of the essential parameters have *uniform a priori distributions*, then the most probable values of the unknown implied by a given set of observational data are given identically by the application of the technique of Least Sum of Squared Residuals. He did not publish these results, however, until 1809, in Book II, Section 3, of his *Theory of the Motion of Heavenly Bodies Moving about the Sun in Sections* (21).

Gauss was well aware that this derivation of his now famous law of error and consequent justification of the technique of *Least Sum of Squared Residuals* was merely an extension of the Principles of the Arithmetic Mean and stood or fell with this Principle. Thus, he remarked that the principle that "the most probable system of values of the unknown quantities [is that for which] the sum of the squares of the differences between the observed and computed values of the functions [observed] is a minimum . . . must, everywhere be considered an axiom with the same propriety as the arithmetical mean of several observed values of the same quantity is adopted as the most probable value" (21, art. 179). But his analysis of the Method of Least Squares remains notable because he recognized that "the constant h can be considered as a measure of the precision [*praecisionis*] of the observations" and then went on to give (1) the formula for the precision of a linear function independent observations of equal or unequal precisions, and (2) the rule for weighting results of unequal precision so as to obtain the combined result of maximum attainable precision. These are everlasting accomplishments of his first 'proof'.

Laplace greatly strengthened Gauss's first 'proof' almost immediately after its publication, by his discovery (22 pp. 383-389) that, under certain very general conditions (not considered in full generality by Laplace) the distributions of linear functions, and hence of the arithmetic means, of n independent errors can be approximated (when properly scaled) by

Gauss's law of error (5), with the error of the approximation tending to zero as $n \to \infty$. From this it follows directly that the Method of Least Squares as developed by Gauss leads to 'most probable values' (under "very general conditions") when the number of independent observations involved is large. The Method of Least Squares was, therefore, regarded as firmly established, not merely on grounds of algebraic and arithmetical convenience, but also via the calculus of probabilities—at least when the number of independent observations is large!

IV. Minimum Errors of Estimation and Gauss's Second 'Proof'

As noted above, Laplace suggested in 1774 (20) that the 'best mean' to take in practical astronomy is that function of the observations which has an equal probability of over- and under-estimating the true value, showed that this is equivalent to adopting the principle of *Least Mean Absolute Error of Estimation*, and gave an algorithm for finding this particular function of three observations in a one-parameter case. By this algorithm his 'best mean' is given by the abscissa $T(y_1, y_2, y_3)$ *that divides the area* under the curve $f(y_1-\tau)f(y_2-\tau)f(y_3-\tau)$, considered as a function of τ, into two equal halves, $f(x)$ being the law of error involved. In 1778 (23), Laplace extended this agreement to the case of n independent observations and termed this procedure "the most advantageous method" of estimation. This approach was invented anew and fully explored by E. J. G. Pitman in 1939 (24). Unfortunately, it usually leads to intractable equations for the "most advantageous" estimates, except for very special choices of the law of error. Thus, in 1811 (25), Laplace found that, among all laws of error of the form $f(x) = Ke^{-\psi(x^2)}$ where $\psi(x^2)$ is an arbitrary continuous function of $x^2 = (y-\tau)^2$, the Gaussian law (5) is the only one for which the arithmetic mean $\bar{Y}$ of n independent observations is the "most advantageous" estimator of τ.

By adopting instead the principle of *Least Mean Squared Error of Estimation* and the requirement that the resulting "best mean" should yield the true values of the quantities concerned if it should happen that all of the observations were entirely free from error, Gauss showed in 1821-23 (26, 27) that, when the resulting 'best values' are linear functions of the observations, then they are identically the same as those given by the technique of *Least Sum of Squared Residuals* (which provides the practical *modus operandi* for obtaining them), and that in this important case the *Least Mean Squared Error* property is completely independent of the law of error involved. This fact, which mathematical statisticians today express by saying that the Method of Least Squares yields *minimum variance linear unbiased estimators* of the unknown magnitudes concerned under "very general conditions", is considered by many mathematical statisticians today to be the *real* theoretical basis of the Method of Least Squares. Henri Poincaré (1854-1912) remarked in 1893-94 (28, p. 168), "This approach justifies the [Method of Least Squares] *independently of the law of errors* . . . . is, thus, a refutation of Gauss's [earlier] reasoning [and] it is rather strange that this refutation is due to Gauss himself". And it is equally surprising that this best-linear-unbiased-estimator property of Least Squares seems to be unknown to many users of the Method of Least Squares today.

V. Concluding Remarks

The robust survival of the Method of Least Squares as a valuable tool of applied science no doubt stems in part from the algebraic and arithmetical advantages of *Least Sum of Squared Residuals* and in part from the fact this procedure also yields estimates of *Least Mean Squared Error* in the important case when the end results are linear functions of the basic observations. This one-to-one correspond-

ence between minimizing some function of the *residuals* and minimizing *the same function* of *Errors of Estimation* appears to be a unique property of *Least Squares*. And although the Method of Least Squares does not lead to the best available estimates of unknown parameters when the law of error is other than the Gaussian, if the number of independent observations available is much larger than the number of parameters to be determined the Method of Least Squares can be usually counted on to yield nearly-best estimates.

References

(1) Russell, Bertrand. *Human Knowledge: Its Scope and Limits*. Simon and Schuster, New York, 1948, p. 344.

(2) Whittaker, E. T. and Robinson, G. *The Calculus of Observations*. Blackie & Son, Ltd., London, 1924. 2nd edition, 1932; 3rd edition, 1940.

(3) Legendre, Adrien Marie. *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris, 1805. Appendix, 'Sur laMéthode des moindres quarrés', pp. 72-80. (English translation of pp. 72-75, by Henry A. Ruger and Helen M. Walker, in David Eugene Smith, *A Source Book in Mathematics*, McGraw-Hill Co., Inc., New York, 1929, pp. 576-579. Reprinted in 2 vols., Dover Publications, Inc., New York, 1959.)

(4) Cotes, Roger. Aestimatio errorum in mixta mathesi, per variationes partium trianguli plani et spherici, *Opera Miscellania* (appended to his *Harmonia Mensurarum*, Cantabrigiae, 1722), pp. 1-22.

(5) Euler, Leonhard. Pièce qui a remporté le prix de l'Academie royale des sciences en 1748, sur les inégalités du mouvement de Saturne et de Jupiter. Paris, 1749.

(6) Mayer, Johann Tobias. Abhandlung über die Umwälzung des Mondes um seine Axe, *Kosmographische Nachrichten und Sammlungen*, Vol. I (1748), pp. 52-183 (published 1750).

(7) Boscovich, Roger Joseph. De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa, *Bononiensi Scientiarum et Artum Instituto Atque Academia Commentarii*, Tomus IV, pp. 353-96, 1757.

(8) Stay, Benedict. *Philosophiae Recentioris, a Benedicto Stay in Romano Archigynasis Publico Eloquentare Professore, versibus traditae, Libri X, cum adnotationibus et Supplementis P. Rogerii Josephi Boscovich* S.J., Tomus II, Romae, 1760.

(9) Eisenhart, Churchill. Boscovich and the Combination of Observations. Chapter 7 in R. J. Boscovich, F.R.S.: *Studies of His Life and Work*, edited by Lancelot Law Whyte, Allen and Unwin, Ltd., London, 1961.

(10) Simon, Pierre, Marquis de Laplace. Mémoire sur la Figure de la Terre. *Mémoires de l' Académie royale des Sciences de Paris, pour l'année 1783*, pp. 17-46, Paris, 1786. Reprinted in *Oeuvres de Laplace*, National Edition, Vol. 11, Gauthier-Villars, Paris, 1895.

(11) Laplace. Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pendule, *Histoire de l'Académie royale des inscriptions et belles lettres, avec les Mémoires de littérature tirées des registres de cette académie, Année* 1789, 18-43 of the Memoires. Paris, 1792.

(12) Laplace. Traité de Mécanique Céleste, Vol. II. Paris, 1799. Reprinted as Vol. II of *Oeuvres de Laplace*, Paris, 1843; National Edition, Gauthier-Villars, Paris, 1878; English translation, with notes and commentary, by Nathaniel Bowditch, Boston, 1832.

(13) Dunnington, Guy Waldo. *Carl Friedrich Gauss: Titan of Science*. Hafner Publishing Co., New York, 1955.

(14) Gauss, Carl Friedrich. Summary of his paper "Theoria combinationis observationum erroribus minimis obnoxiae, pars prior." *Göttingische gelehrte Anzeigen*, February 26, 1821; Gauss *Werke*, vol. IV, p. 96-100.

(15) Simpson, Thomas. A letter to the Right Honourable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations, in practical astronomy, *Phil. Trans. Roy. Soc. London*, Vol. 49, Part I, pp. 82-93, 1755.

(16) Demoivre, Abraham. *The Doctrine of Chances: or a Method of Calculating the Probability of Events in Play*. London, 1718 (2nd ed., 1738; 3rd ed., 1756).

(17) Simpson. An attempt to show the advantage arising by taking the mean of a number of observations in practical astronomy. In his *Miscellaneous Tracts on Some Curious and Very Interesting Subjects in Mechanics, Physical-Astronomy, and Speculative Mathematics*, pp. 64-75. J. Nourse, London, 1757.

(18) Joseph-Louis, Comte de Lagrange. Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observations; dans lequel on examine les avantages de cette méthode par le calcul des probabilités; et où l'on résoud différents problèmes relatifs à cette matière. *Miscellanea Taurinensia*, Vol. 5 (1770-1773), pp. 167-232 of Mathematics portion. Reprinted in *Oeuvres de Lagrange*, Vol. 2, pp. 173-234. Gauthier-Villars, Paris, 1868.

(19) Bernoulli, Daniel. Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda. *Acta Academiae Scientiorum Petropolitanae*, Vol. I (1777), Pt. I, pp. 3-23 of the Memoirs. St.

Petersburg, 1778; English translation by C. G. Allen, *Biometrika*, Vol. 48, Pts. 1 and 2 (June 1961), pp. 3-13.

(20) Laplace. Problème III: Déterminer le milieu que l'on doit prendre entre trois observations données d'un même phénomène. Pp. 634-644 of his "Mémoire sur la probabilité des causes par les événements", *Mémoires de Mathématique et de Physique*, Vol. 6, pp. 621-657, Paris, 1774.

(21) Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Frid. Perthes et I. H. Besser, Hamburg, 1809; Reprinted in *Carl Friedrich Gauss Werke*, Band VII. Gotha, 1871. English translation by Charles Henry Davis, Boston, 1857; reprinted by Dover Publications, Inc., New York, 1963.

(22) Laplace. Mémoire sur les approximations des formu.es qui sont fonctions de trèsgrands nombres, et sur leur application aux probabilités, *Mémoires de la classe des Sciences Mathématiques et Physiques, de l'Institut de France Année 1809*, pp. 353-415; Supplement pp. 559-565. Paris, 1810.

(23) Laplace. Mémoire sur les Probabilités, *Histoire de l'Académie royale des Sciences de Paris*, Annee 1778, pp. 227-332. Paris, 1781.

(24) Pitman, E. J. G. The estimation of the location and scale parameters of a continuous population of any given form, *Biometrika*, Vol. XXX, Parts 3 and 4 (Jan. 1939), pp. 391-421.

(25) Laplace. Mémoire sur les intégrales définies et leur application aux probabilités, et spécialement à la recherche du milieu qu'il faut choisir entre les résultats des observations, *Mémoires de la Classe des Sciences Mathématiques et Physiques de l'Institut Impérial de France, Année 1810*, pp. 279-347. Paris, 1811.

(26) Gauss. Theoria combinationis observationum erroribus minimis obnoxiae. Pars prior. [presented 15 Feb. 1821.] *Commentationes societas regiae scientiarum Göttingensis recentiores*, Vol. V, pp. 33-62, 1823. Reprinted in *Carl Friedrich Gauss Werke*, Vol. IV, pp. 1-26. Göttingen, 1873.

(27) Gauss. Theoria combinationis . . . Pars posterior. [Presented 2 Feb. 1823.] *Commentationes . . .*, Vol. V, pp. 63-90, 1823. Reprinted in *Werke*, Vol. IV, pp. 29-53. Göttingen, 1873.

(28) Henri Poincaré. *Calcul des probabilités. Leçons professées pendant le deuxième semestre 1893-1894*. Georges Carré, Paris, 1896.

5. Statistical Treatment of Measurement Data

Foreword

Since there are available many excellent textbooks on statistical methods, we include in this section only those entries which are either (1) addressed specifically to metrologists, or (2) topics not fully discussed in other texts.

Two introductory treatments of statistical concepts and terminology are provided. Some Basic Statistical Concepts and Preliminary Considerations by Mary G. Natrella and Churchill Eisenhart (5.1) introduces the fundamental ideas of *populations*, *samples*, and *distributions* that underlie all statistical procedures. Then it discusses the interpretation of some important types of procedures — estimation, confidence intervals, tolerance intervals, and statistical tests of significance.

Statistical Concepts in Metrology by Harry H. Ku (5.2) deals with basic statistical concepts as applied to the description and characterization of a measurement process. A basic kit of tools for the manipulation of measurement data is given, and their use for evaluation of precision is discussed. The use of control chart techniques for monitoring stability is emphasized.

The excerpt, Randomization in Factorial and Other Experiments (5.4), from E. B. Wilson's Introduction to Scientific Research contains two examples illustrating the importance of randomization in experiments. We have included it to add emphasis to the point that the experimenter cannot take for granted that his data conform to the conditions (expressed as assumptions) underlying proper use of statistical techniques.

The use of propagation of error formulas has had a long history; yet the term does not appear often in the index of current statistical textbooks. Ku's Notes (5.3) differentiates the two types of usage and outlines conditions under which these formulas give good approximations.

Detection and rejection of outliers is a problem that forever plagues the experimenter. Kruskal's remarks (5.5) gather together in one place some nontechnical thoughts on this matter, whereas Proschan's paper (5.6) lists several operational criteria. For further reading we suggest two papers, one by F. J. Anscombe and one by C. Daniel, in Technometrics, vol. 2, no. 2, May, 1960.

CHAPTER 1 *

SOME BASIC STATISTICAL CONCEPTS AND PRELIMINARY CONSIDERATIONS

Mary G. Natrella and Churchill Eisenhart

1-1 INTRODUCTION

Statistics deals with the collection, analysis, interpretation, and presentation of numerical data. Statistical methods may be divided into two classes—descriptive and inductive. Descriptive statistical methods are those which are used to summarize or describe data. They are the kind we see used everyday in the newspapers and magazines. Inductive statistical methods are used when we wish to generalize from a small body of data to a larger system of similar data. The generalizations usually are in the form of estimates or predictions. In this handbook we are mainly concerned with inductive statistical methods.

1-2 POPULATIONS, SAMPLES, AND DISTRIBUTIONS

The concepts of a *population* and a *sample* are basic to inductive statistical methods. Equally important is the concept of a *distribution.*

Any finite or infinite collection of individual things—objects or events—constitutes a *population.* A population (also known as a universe) is thought of not as just a heap of things specified by enumerating them one after another, but rather as an aggregate determined by some property that distinguishes between things that do and things that do not belong. Thus, the term *population* carries with it the connotation of completeness. In contrast, a *sample* defined as a portion of a population, has the connotation of incompleteness.

Examples of populations are:

(a) The corporals in the Marines on July 1, 1956.

(b) A production lot of fuzes.

(c) The rounds of ammunition produced by a particular production process.

(d) Fridays the 13th.

(e) Repeated weighings of the powder charge of a particular round of ammunition.

(f) Firings of rounds from a given production lot.

In examples (a), (b), and (c), the "individuals" comprising the population are material objects (corporals, fuzes, rounds); in (d) they are periods of time of a very restricted type; and in (e) and (f) they are physical operations. Populations (a) and (b) are clearly finite, and their constituents are determined by the official records of the Marine Corps and the appropriate production records, respectively. Populations (c), (d), and (e) are conceptually infinite. Offhand, the population example (f) would

* NBS Handbook 91, 1966.

seem to be finite, because firing is a destructive operation; but in order to allow for variation in quality among "firings" performed in accordance with the same general procedure it is sometimes useful, by analogy with repetitive weighings, to regard an actual firing as a sample of size one from a conceptually infinite population of "possible" firings, any one of which might have been associated with the particular round conceived. In this connection, note that in examples (e) and (f) the populations involved are not completely defined until the weighing and firing procedures concerned have been fully specified.

Attention to some characteristic of the individuals of a population that is not the same for every individual leads immediately to recognition of the *distribution* of this characteristic in the population. Thus, the heights of the corporals in the Marines on July 1, 1956, the burning times of a production lot of fuzes, and the outcomes of successive weighings of a powder charge ("observed weights" of the charge) are examples of distributions. The presence or absence of an attribute is a characteristic of an individual in a population, such as "tatooed" or "not tatooed" for the privates in the Marines. This kind of characteristic has a particularly simple type of distribution in the population.

Attention to one, two, three, or more characteristics for each individual leads to a univariate, bivariate, trivariate, or multivariate distribution in the population. The examples of populations given previously were examples of univariate distributions. Simultaneous consideration of the muzzle velocities and weights of powder charges of rounds of ammunition from a given production process determines a bivariate distribution of these characteristics in the population. Simultaneous recognition of the frequencies of each of a variety of different types of accidents on Friday the 13th leads to a multivariate distribution. In connection with these examples, note that, as a general principle, the distribution of a characteristic or a group of characteristics in a population is not completely defined until the method or methods of measurement or enumeration involved are fully specified.

The distribution of some particular property of the individuals in a population is a collective property of the population; and so, also, are the average and other characteristics of the distribution. The methods of inductive statistics enable us to learn about such population characteristics from a study of samples.

An example will illustrate an important class of derived distributions. Suppose we select 10 rounds of ammunition from a given lot and measure their muzzle velocities when the rounds are fired in a given test weapon. Let $\bar{X}$ be the average muzzle velocity of the 10 rounds. If the lot is large, there will be many different sets of 10 rounds which could have been obtained from the lot. For each such sample of 10 rounds, there will correspond an average muzzle velocity $\bar{X}_i$. These averages, from all possible samples of 10, themselves form a distribution of sample averages. This kind of distribution is called the *sampling distribution of* $\bar{X}$ *for samples of size 10* from the population concerned. Similarly, we may determine the *range* R of muzzle velocities (i.e., the difference between the largest and the smallest) for each of all possible samples of 10 rounds each. These ranges R_i ($i = 1, 2, \ldots$) collectively determine the *sampling distribution of the range* of muzzle velocities *in samples of size 10* from the population concerned. The methods of inductive statistics are based upon the mathematical properties of sampling distributions of sample *statistics* such as $\bar{X}$ and R.

Let us summarize: A population in Statistics corresponds to what in Logic is termed the "universe of discourse"—it's what we are talking about. By the methods of inductive statistics we can learn, from a study

of samples, only about population characteristics—only about *collective* properties of the populations represented by the individuals in the samples—not about characteristics of specific individuals with unique idiosyncrasies. The population studied may be large or small, but there must be a population; and it should be well defined. The characteristic of interest must be a collective property of the population.

1-3 STATISTICAL INFERENCES AND SAMPLING

1-3.1 STATISTICAL INFERENCES

If we were willing or able to examine an entire population, our task would be merely that of describing that population, using whatever numbers, figures, or charts we cared to use. Since it is ordinarily inconvenient or impossible to observe every item in the population, we take a sample—a portion of the population. Our task is now to generalize from our observations on this portion (which usually is small) to the population. Such generalizations about characteristics of a population from a study of one or more samples from the population are termed *statistical inferences.*

Statistical inferences take two forms: *estimates* of the magnitudes of population characteristics, and *tests of hypotheses* regarding population characteristics. Both are useful for determining which among two or more courses of action to follow in practice when the "correct" course is determined by some particular but unknown characteristic of the population.

Statistical inferences all involve reaching conclusions about population characteristics (or at least acting as if one had reached such conclusions) from a study of samples which are known or assumed to be portions of the population concerned. Statistical inferences are basically predictions of what would be found to be the case if the parent populations could be and were fully analyzed with respect to the relevant characteristic or characteristics.

A simple example will serve to bring out a number of essential features of statistical inferences and the methods of inductive statistics. Suppose that four cards have been drawn from a deck of cards and have been found to be the Ace of Hearts, the Five of Diamonds, the Three of Clubs, and the Jack of Clubs. The specific methods discussed in the following paragraphs will be illustrated from this example.

First of all, from the example, we can clearly conclude at once that the deck contained at least one Heart, at least one Diamond, and at least two Clubs. We also can conclude from the presence of the Five and the Three that the deck is definitely not a pinochle deck. These are perhaps trivial inferences, but their validity is above question and does not depend in any way on the *modus operandi* of drawing the four cards.

In order to be able to make inferences of a more substantial character, we must know the nature of the sampling operation that yielded the sample of four cards actually obtained. Suppose, for example, that the sampling procedure was as follows: The cards were drawn in the order listed, each card being selected *at random* from all the cards present in the deck when the card was drawn. This defines a hypothetical population of drawings. By using an appropriate technique of inductive statistics—essentially, a "catalog" of all possible samples of four, showing for each sample the conclusion to be adopted whenever that sample occurs—we can make statistical inferences about properties of this population of drawings. The statistical inferences made will be rigorous if, and only if, the inductive technique

used is appropriate to the sampling procedure actually employed.

Thus, by taking the observed proportion of Clubs as an estimate of the proportion of Clubs in the abstract population of drawings, we may assert: the proportion of Clubs is 50%. Since random sampling of the type assumed assures that the proportion of Clubs in the population of drawings is the same as the proportion of Clubs in the deck, we may assert with equal validity: the proportion of Clubs in the deck is 50%. If the deck concerned actually was a standard bridge deck, then in the present instance our estimate is wrong in spite of being the best single estimate available.

We know from experience that with samples of four we cannot expect to "hit the nail on the head" every time. If instead of attempting to make a single-number estimate we had chosen to refer to a "catalog" of *interval estimates* (see, for example, Table A-22*), we would have concluded that the proportion of Clubs is between 14% and 86% inclusive, with an expectation of being correct 9 times out of 10. If the deck was in fact a standard bridge deck, then our conclusion is correct in this instance, but its validity depends on whether the sampling procedure employed in drawing the four cards corresponds to the sampling procedure assumed in the preparation of the "catalog" of answers.

It is important to notice, moreover, that strictly we have a right to make statistical inferences only with respect to the hypothetical population of drawings defined by the sampling operation concerned. In the present instance, as we shall see, the sampling operation was so chosen that the parameters (i.e., the proportions of Hearts, Clubs, and Diamonds) of the hypothetical population of drawings coincide with the corresponding parameters of the deck. Hence, in the present case, inferences about the parameters of the population of drawings may be interpreted as inferences about the composition of the deck. This emphasizes the importance of selecting and employing a sampling procedure such that the relevant parameters of the population of drawings bear a known relation to the corresponding parameters of the real-life situation. Otherwise, statistical inferences with respect to the population of drawings carried over to the real-life population will be lacking in rigor, even though by luck they may sometimes be correct.

1-3.2 RANDOM SAMPLING

In order to make valid nontrivial generalizations from samples about characteristics of the populations from which they came, the samples must have been obtained by a sampling scheme which insures two conditions:

(a) Relevant characteristics of the populations sampled must bear a known relation to the corresponding characteristics of the population of all possible samples associated with the sampling scheme.

(b) Generalizations may be drawn from such samples in accordance with a given "book of rules" whose validity rests on the mathematical theory of probability.

If a sampling scheme is to meet these two requirements, it is necessary that the selection of the individuals to be included in a sample involve some type of *random selection*, that is, each possible sample must have a fixed and determinate probability of selection. (For a very readable expository discussion of the general principles of sampling, with examples of some of the more common procedures, see the article by Cochran, Mosteller, and Tukey[1]. For fuller details see, for example, Cochran's book[2].

The most widely useful type of random selection is *simple* (or *unrestricted*) *random sampling*. This type of sampling is defined by the requirement that each individual in the population has an equal chance of being the first member of the sample; after the

* The A-Tables referenced in this handbook are contained in Section 5, ORDP 20-114.

first member is selected, each of the remaining individuals in the population has an equal chance of being the second member of the sample; and so forth. For a sampling scheme to qualify as simple random sampling, it is not sufficient that "each individual in the population have an equal chance of appearing in the sample," as is sometimes said, but it is sufficient that "each possible sample have an equal chance of being selected." Throughout this handbook, we shall assume that all samples are random samples in the sense of having been obtained by simple random sampling.

It cannot be overemphasized that the *randomness* of a sample is inherent in the sampling scheme employed to obtain the sample and not an intrinsic property of the sample itself. Experience teaches that it is not safe to assume that a sample selected haphazardly, without any conscious plan, can be regarded as if it had been obtained by simple random sampling. Nor does it seem to be possible to consciously draw a sample *at random*. As stated by Cochran, Mosteller, and Tukey[1],

> We insist on some semblance of mechanical (dice, coins, random number tables, etc.) randomization before we treat a sample from an existent population as if it were random. We realize that if someone just "grabs a handful," the individuals in the handful almost always resemble one another (on the average) more than do the members of a simple random sample. Even if the "grabs" are randomly spread around so that every individual has an equal chance of entering the sample, there are difficulties. Since the individuals of grab samples resemble one another *more* than do individuals of random samples, it follows (by a simple mathematical argument) that the means of grab samples resemble one another *less* than the means of random samples of the same size. From a grab sample, therefore, we tend to *under*estimate the variability in the population, although we should have to *over*estimate it in order to obtain valid estimates of variability of grab sample means by substituting such an estimate into the formula for the variability of means of simple random samples. Thus, using simple random sample formulas for grab sample means introduces a double bias, both parts of which lead to an unwarranted appearance of higher stability.

Instructions for formally drawing a sample at random from a particular population are given in Paragraph 1-4.

Finally, it needs to be noticed that a particular sample often qualifies as "a sample" from any one of several populations. For example, a sample of n rounds from a single carton is a sample from that carton, from the production lot of which the rounds in that carton are a portion, and from the production process concerned. By drawing these rounds from the carton in accordance with a simple random sampling scheme, we can insure that they are a (simple) random sample from the carton, not from the production lot or the production process. Only if the production process is in a "state of statistical control" may our sample also be considered to be a simple random sample from the production lot and the production process. In a similar fashion, a sample of repeated weighings can validly be considered to be a random sample from the conceptually infinite population of repeated weighings by the same procedure only if the weighing procedure is in a state of statistical control (see Chapter 18, in Section 4, ORDP 20-113).

It is therefore important in practice to know from which of several possible "parent" populations a sample was obtained *by simple random sampling*. This population is termed the *sampled population*, and may be quite different from the population of interest, termed the *target population*, to which we would like our conclusions to be applicable. In practice, they are rarely identical, though the difference is often small. A sample from the target population of rounds of ammunition produced by a particular production process will actually be a sample from one or more production lots (sampled population), and the difference between sampled and target populations will be smaller if the sampled population comprises a larger number of production lots. The further the sampled population is removed from the target population, the more the burden of validity of conclusions is shifted from the shoulders of the statistician to those of the subject matter expert, who must place greater and greater (and perhaps unwarranted) reliance on "other considerations."

1-4 SELECTION OF A RANDOM SAMPLE

As has been brought out previously, the method of choosing a sample is an all-important factor in determining what use can be made of it. In order for the techniques described in this handbook to be valid as bases for making statements from samples about populations, we must have unrestricted random samples from these populations. In practice, it is not always easy to obtain a random sample from a given population. Unconscious selections and biases tend to enter. For this reason, it is advisable to use a table of random numbers as an aid in selecting the sample. Two tables of random numbers which are recommended are by L. H. C. Tippett[3] and The Rand Corporation[4]. These tables contain detailed instructions for their use. An excerpt from one of these tables[4] is given in Table A-36. This sample is included for illustration only; a larger table should be used in any actual problem. Repeated use of the same portion of a table of random numbers will not satisfy the requirements of randomness.

An illustration of the method of use of tables of random numbers follows. Suppose the population consists of 87 items, and we wish to select a random sample of 10. Assign to each individual a separate two-digit number between 00 and 86. In a table of random numbers, pick an arbitrary starting place and decide upon the direction of reading the numbers. Any direction may be used, provided the rule is fixed in advance and is independent of the numbers occurring. Read two-digit numbers from the table, and select for the sample those individuals whose numbers occur until 10 individuals have been selected. For example, in Table A-36, start with the second page of the Table (p. T-83), column 20, line 6, and read down. The 10 items picked for the sample would thus be numbers 38, 44, 13, 73, 39, 41, 35, 07, 14, and 47.

The method described is applicable for obtaining simple random samples from any sampled population consisting of a finite set of individuals. In the case of an infinite sampled population, these procedures do not apply. Thus, we might think of the sampled population for the target population of weighings as comprising all weighings which might conceptually have been made during the time while weighing was done. We cannot by mechanical randomization draw a random sample from this population, and so must recognize that we have a random sample only *by assumption.* This assumption will be warranted if previous data indicate that the weighing procedure is in a state of statistical control; unwarranted if the contrary is indicated; and a leap in the dark if no previous data are available.

1-5 SOME PROPERTIES OF DISTRIBUTIONS

Although it is unusual to examine populations in their entirety, the examination of a large sample or of many small samples from a population can give us much information about the general nature of the population's characteristics.

One device for revealing the general nature of a population distribution is a histogram. Suppose we have a large number of observed items and a numerical measurement for each item, such as, for example, a Rockwell hardness reading for each of 5,000 specimens. We first make a table showing the numerical measurement and the number of times (i.e., frequency) this measurement was recorded.

Rockwell Hardness Number	Frequency
55	1
56	17
57	135
58	503
59	1,110
60	1,470
61	1,120
62	490
63	125
64	26
65	3

Data taken, by permission, from *Sampling Inspection by Variables* by A. H. Bowker and H. P. Goode, Copyright, 1952, McGraw-Hill Book Company, Inc.

From this frequency table we can make the histogram as shown in Figure 1-1. The height of the rectangle for any hardness range is determined by the number of items in that hardness range. The rectangle is centered at the tabulated hardness value. If we take the sum of all the rectangular areas to be one square unit, then the area of an individual rectangle is equal to the *proportion* of items in the sample that have hardness values in the corresponding range. When the sample is large, as in the present instance, the histogram may be taken to exemplify the general nature of the corresponding distribution in the population.

If it were possible to measure hardness in finer intervals, we would be able to draw a larger number of rectangles, smaller in width than before. For a sufficiently large sample and a sufficiently fine "mesh," we would be justified in blending the tops of the rectangles into a continuous curve, such as that shown in Figure 1-2, which we could expect to more nearly represent the underlying population distribution.

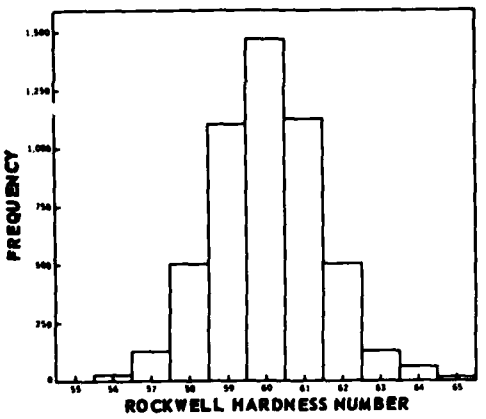


Figure 1-1. Histogram representing the distribution of 5,000 Rockwell hardness readings.

Reproduced by permission from *Sampling Inspection by Variables* by A. H. Bowker and H. P. Goode, Copyright, 1952, McGraw-Hill Book Company, Inc.

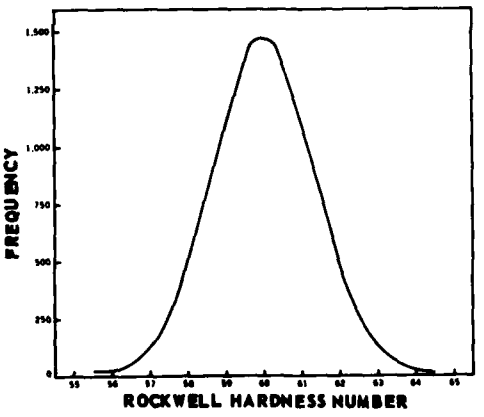


Figure 1-2. Normal curve fitted to the distribution of 5,000 Rockwell hardness readings.

Reproduced by permission from *Sampling Inspection by Variables* by A. H. Bowker and H. P. Goode, Copyright, 1952, McGraw-Hill Book Company, Inc.

If we were to carry out this sort of scheme on a large number of populations, we would find that many different curves would arise, as illustrated in Figure 1-3. Possibly, the majority of them would resemble the class of symmetrical bell-shaped curves called "normal" or "Gaussian" distributions, an example of which is shown in the center of Figure 1-3. A normal distribution is unimodal, i.e., has only a single highest point or *mode*, as also are the two asymmetrical curves in the lower left and upper right of Figure 1-3.

A "normal" distribution is completely determined by two parameters: m, the arithmetic mean (or simply "the mean") of the distribution, and σ, the standard deviation (often termed the "population mean" and "population standard deviation"). The *variance* of the distribution is σ^2. Since a normal curve is both unimodal and symmetrical,

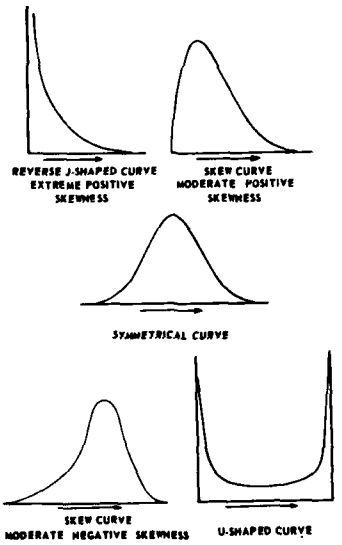


Figure 1-3. Frequency distributions of various shapes.

Adapted with permission from *Elements of Statistical Reasoning* by A. E. Treloar, Copyright, 1939, John Wiley & Sons, Inc.

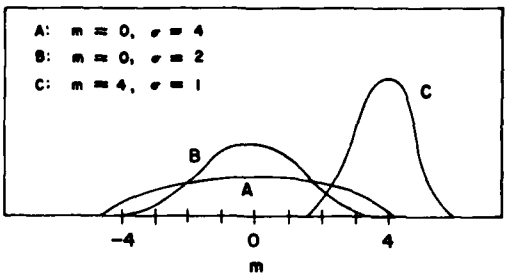


Figure 1-4. Three different normal distributions.

m is also the *mode* and the value which divides the area under the curve in half, i.e., the *median*. It is useful to remember that σ is the distance from m to either of the two inflection points on the curve. (The inflection point is the point at which the curve changes from concave upward to concave downward.) This is a special property of the normal distribution. More generally, the mean of a distribution m is the "center of gravity" of the distribution; σ is the "radius of gyration" of the distribution about m, in the language of mechanics; and σ^2 is the second moment about m.

The parameter m is the *location parameter* of a normal distribution, while σ is a measure of its spread, scatter, or dispersion. Thus, a change in m merely slides the curve right or left without changing its profile, while a change in σ widens or narrows the curve without changing the location of its center. Three different normal curves are shown in Figure 1-4. (All normal curves in this section are drawn so that the area under the curve is equal to one, which is a standard convention.)

Figure 1-5 shows the percentage of elements of the population contained in various intervals of a normal distribution. z is the distance from the population mean in units of the standard deviation and is computed using the formula $z = (X-m)/\sigma$, where X represents any value in the population. Using z to enter Table A-1, we find P, the proportion

of elements in the population which have values of z smaller than any given z. Thus, as shown in Fig. 1-5, 34.13% of the population will have values of z between 0 and 1 (or between 0 and −1); 13.59% of the population, between 1 and 2 (or between −1 and −2); 2.14% between 2 and 3 (or between −2 and −3); and .14% beyond 3 (or beyond −3). Figure 1-5 shows these percentages of the population in various intervals of z.

For example, suppose we know that the chamber pressures of a lot of ammunition may be represented by a normal distribution, with the average chamber pressure m = 50,000 psi and standard deviation σ = 5,000 psi. Then $z = \frac{X-50,000}{5,000}$ and we know (Fig. 1-5) that if we fired the lot of ammunition in the prescribed manner we would expect 50% of the rounds to have a chamber pressure above 50,000 psi, 15.9% to have pressures above 55,000 psi, and 2.3% to have pressures above 60,000 psi, etc.

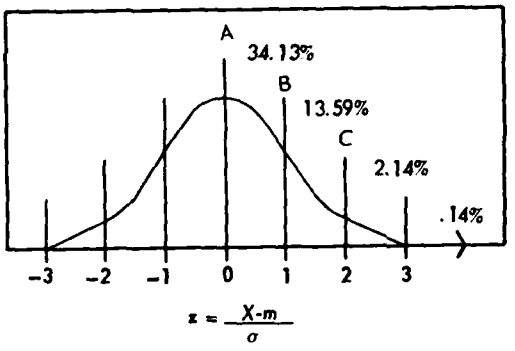


Figure 1-5. Percentage of the population in various intervals of a normal distribution.

1-6 ESTIMATION OF m and σ

In areas where a lot of experimental work has been done, it often happens that we know m or σ, or both, fairly accurately. However, in the majority of cases it will be our task to estimate them by means of a sample. Suppose we have n observations, $X_1, X_2, \ldots, X_n$ taken at random from a normal population. From a sample, what are the best estimates of m and σ? Actually, it is usual to compute the best unbiased estimates of m and σ^2, and then take the square root of the estimate of σ^2 as the estimate of σ. These recommended estimates of m and σ^2 are:*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

$\bar{X}$ and s^2 are the *sample mean* and *sample estimate of variance*, respectively. (s is often called "the sample standard deviation," but this is not strictly correct and we shall avoid the expression and simply refer to s.) For computational purposes, the following formula for s^2 is more convenient:

$$s^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$$

Nearly every sample will contain different individuals, and thus the estimates $\bar{X}$ and s^2 of m and σ^2 will differ from sample to sample. However, these estimates are such that "on the average" they tend to be equal to m and σ^2, respectively, and in this sense are *unbiased*. If, for example, we have a large number of random samples of size n, the average of their respective estimates of σ^2 will tend to be near σ^2. Furthermore, the amount of fluctuation of the respective s^2's about σ^2 (or of the $\bar{X}$'s about m, if we are estimating m) will be smaller in a certain well-defined sense than the fluctuation would be for any estimates other than the recommended ones. For these reasons, $\bar{X}$ and s^2 are called the "best unbiased" estimates of m and σ^2, respectively.*

As might be expected, the larger the sample size n, the more faith we can put in the estimates $\bar{X}$ and s^2. This is illustrated in Figures 1-6 and 1-7. Figure 1-6 shows the distribution of $\bar{X}$ (sample mean) for samples of various sizes from the same normal distribution. The curve for $n = 1$ is the distribution for individuals in the population. All of the curves are centered at m, the popula-

* The Greek symbol Σ is often used as shorthand for "the sum of." For example,

$$\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4$$

$$\sum_{i=1}^{3} (X_i + Y_i) = (X_1 + Y_1) + (X_2 + Y_2) + (X_3 + Y_3)$$

$$\sum_{i=1}^{3} X_i Y_i = X_1Y_1 + X_2Y_2 + X_3Y_3$$

$$\sum_{i=1}^{3} c = c + c + c = 3c$$

* On the other hand, s is not an unbiased estimator of σ. Thus, in samples of size n from a normal distribution, the situation is:

Sample size, n	s is an unbiased estimator of:
2	0.797 σ
3	0.886
4	0.921
5	0.940
6	0.952
7	0.959
8	0.965
9	0.969
10	0.973
20	0.987
30	0.991
40	0.994
60	0.996
120	0.998
∞	1.000

tion mean, but the scatter becomes less as n gets larger. Figure 1-7 shows the distribution of s^2 (sample variance) for samples of various sizes from the same normal distribution.

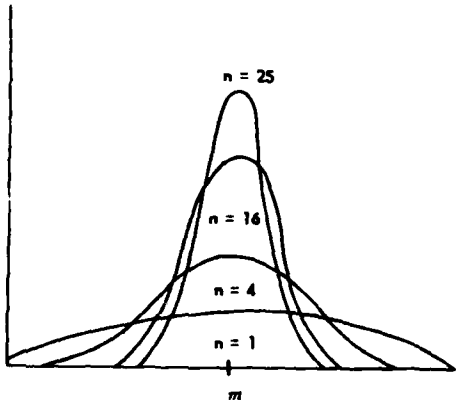


Figure 1-6. Sampling distribution of $\bar{X}$ for random samples of size n from a normal population with mean m.

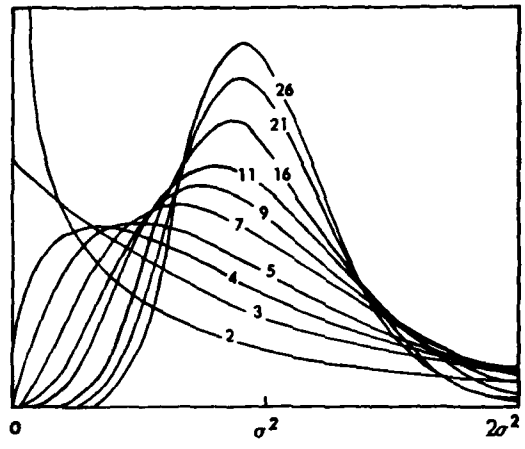


Figure 1-7. Sampling distribution of s^2 for sample size n from a normal population with $\sigma = 1$.

1-7 CONFIDENCE INTERVALS

Inasmuch as estimates of m and σ vary from sample to sample, interval estimates of m and σ may sometimes be preferred to "single-value" estimates. Provided we have a random sample from a normal population, we can make interval estimates of m or σ with a chosen degree of confidence. The level of confidence is not associated with a particular interval, but is associated with the method of calculating the interval. The interval obtained from a particular sample either brackets the true parameter value (m or σ, whichever we are estimating) or does not. The confidence coefficient γ is simply the proportion of samples of size n for which intervals computed by the prescribed method may be expected to bracket m (or σ). Such intervals are known as *confidence intervals*, and always are associated with a prescribed confidence coefficient. As we would expect, larger samples tend to give narrower confidence intervals for the same level of confidence.

Suppose we are given the lot of ammunition mentioned earlier (Par. 1-5) and wish to make a confidence interval estimate of the average chamber pressure of the rounds in the lot. The true average is 50,000 psi,

although this value is unknown to us. Let us take a random sample of four rounds and from this sample, using the given procedure, calculate the upper and lower limits for our confidence interval. Consider all the possible samples of size 4 that could have been taken, and the resulting confidence intervals computed from each. If we compute 50% (90%) confidence intervals, then we expect 50% (90%) of the computed intervals to cover the true value, 50,000 psi. See Figure 1-8.

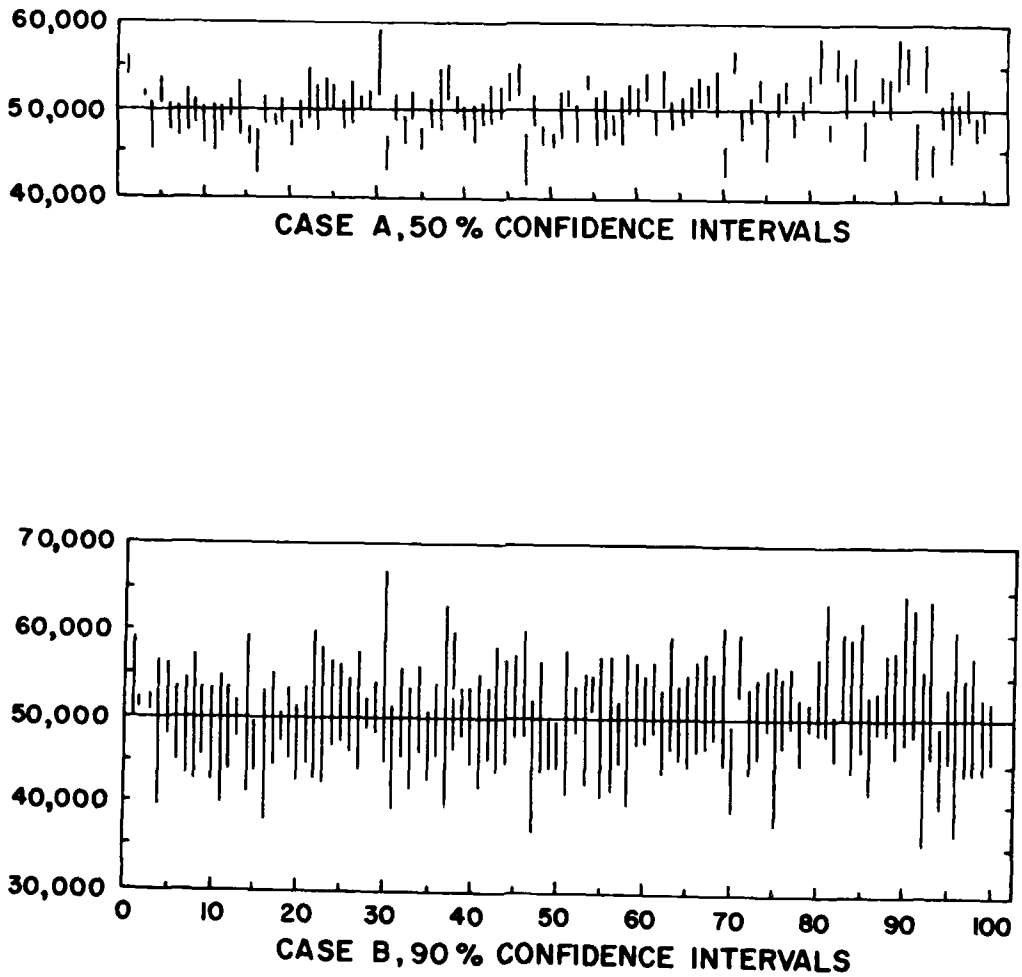


Figure 1-8. Computed confidence intervals for 100 samples of size 4 drawn at random from a normal population with m = 50,000 psi, σ = 5,000 psi. Case A shows 50% confidence intervals; Case B shows 90% confidence intervals.

Adapted with permission from *ASTM Manual on Quality Control of Materials*, Copyright, 1951, American Society for Testing Materials.

In Case A of Figure 1-8, 51 of the 100 intervals actually include the true mean. For 50% confidence interval estimates, we would expect in the long run that 50% of the intervals would include the true mean. Fifty-one out of 100 is a reasonable deviation from the expected 50%. In Case B, 90 out of 100 of the intervals contain the true mean. This is precisely the expected number for 90% intervals.

Note also (Fig. 1-8) that the successive confidence intervals vary both in position and width. This is because they were computed (see Par. 2-1.4) from the sample statistics $\bar{X}$ and s, both of which vary from sample to sample. If, on the other hand, the standard deviation of the population distribution σ were known, and the confidence intervals were computed from the successive $\bar{X}$'s and σ (procedure given in Par. 2-1.5), then the resulting confidence intervals would all be the same width, and would vary in position only.

Finally, as the sample size increases, confidence intervals tend not only to vary less in both position and width, but also to "pinch in" ever closer to the true value of the population parameter concerned, as illustrated in Figure 1-9.

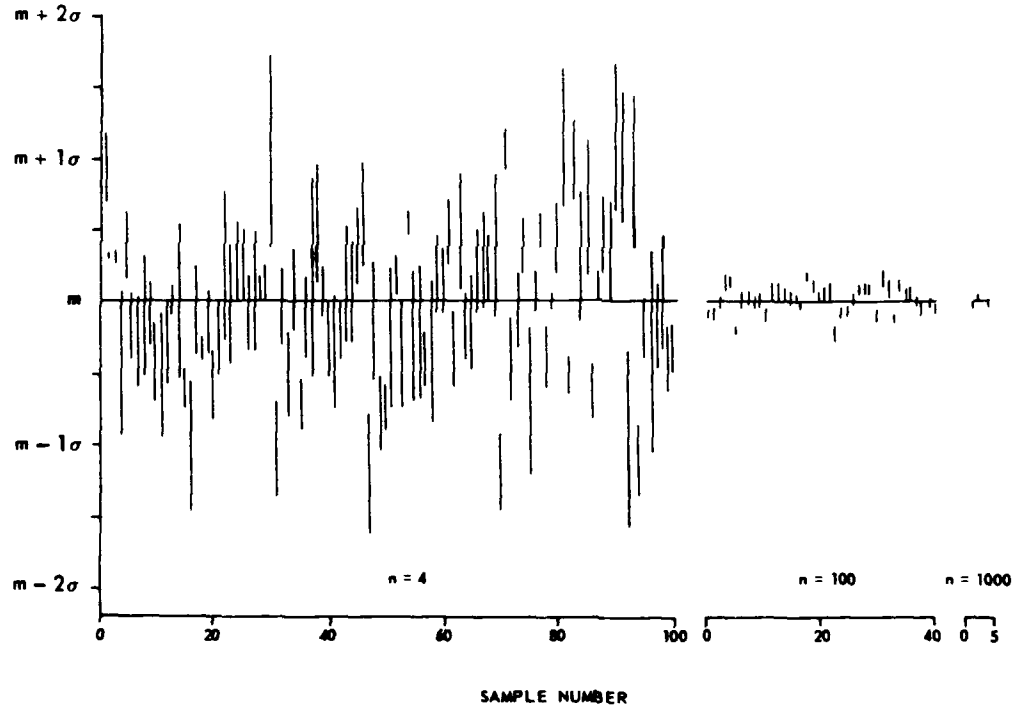


Figure 1-9. Computed 50% confidence intervals for the population mean m from 100 samples of 4, 40 samples of 100, and 4 samples of 1000.

Adapted with permission from *Statistical Method from the Viewpoint of Quality Control* by W. A. Shewhart (edited by W. Edwards Deming), Copyright, 1939, Graduate School, U.S. Department of Agriculture, Washington, D. C.

1-8 STATISTICAL TOLERANCE LIMITS

Sometimes what is wanted is not an estimate of the mean and variance of the population distribution but, instead, two outer values or limits which contain nearly all of the population values. For example, if extremely low chamber pressures or extremely high chamber pressures might cause serious problems, we may wish to know approximate limits to the range of chamber pressures in a lot of ammunition. More specifically, we may wish to know within what limits 99%, for example, of the chamber pressures lie. If we knew the mean m and standard deviation σ of chamber pressures in the lot, and if we knew the distribution of chamber pressures to be normal (or very nearly normal), then we could take $m - 3\sigma$ and $m + 3\sigma$ as our limits, and conclude that approximately 99.7% of the chamber pressures lie within these limits (see Fig. 1-5). If we do not know m and σ, then we may endeavor to approximate the limits with *statistical tolerance limits* of the form $\bar{X} - Ks$ and $\bar{X} + Ks$, based on the sample statistics $\bar{X}$ and s, with K chosen so that we may expect these limits to include *at least* P percent of the chamber pressures in the lot, at some prescribed level of confidence α.

Three sets of such limits for $P = 99.7\%$, corresponding to sample sizes $n = 4$, 100, and 1,000, are shown by the bars in Figure 1-10. It should be noted that for samples of size 4, the bars are very variable both in location and width, but that for $n = 100$ and $n = 1{,}000$, they are of nearly constant width

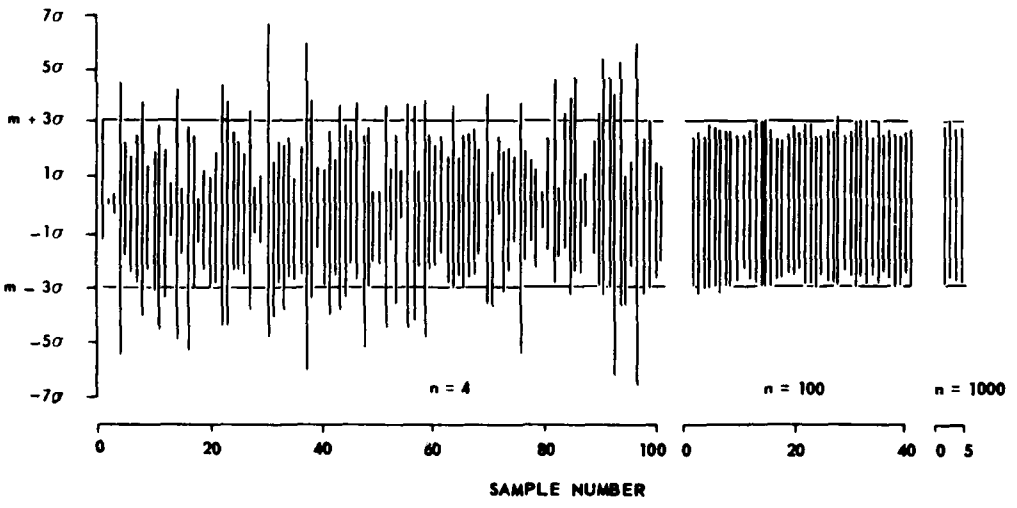


Figure 1-10. Computed statistical tolerance limits for 99.7% of the population from 100 samples of size 4, 40 samples of size 100, and 4 samples of size 1000.

Adapted with permission from *Statistical Method from the Viewpoint of Quality Control* by W. A. Shewhart (edited by W. Edwards Deming). Copyright, 1939, Graduate School, U.S. Department of Agriculture, Washington, D. C.

and position—and their end points approximate very closely to $m - 3\sigma$ and $m + 3\sigma$. In other words, statistical tolerance intervals tend to a fixed size (which depends upon P) as the sample size increases, whereas confidence intervals shrink down towards zero width with increasing sample size, as illustrated in Figure 1-9.

The difference in the meanings of the terms *confidence intervals, statistical tolerance limits,* and *engineering tolerance limits* should be noted. A *confidence interval* is an interval within which we estimate a given population parameter to lie (e.g., the population mean m with respect to some characteristic). *Statistical tolerance limits* for a given population are limits within which we expect a stated proportion of the population to lie with respect to some measurable characteristic. *Engineering tolerance limits* are specified outer limits of acceptability with respect to some characteristic usually prescribed by a design engineer.

1-9 USING STATISTICS TO MAKE DECISIONS

1-9.1 APPROACH TO A DECISION PROBLEM

Consider the following more-or-less typical practical situation: Ten rounds of a new type of shell are fired into a target, and the depth of penetration is measured for each round. The depths of penetration are 10.0, 11.1, 10.5, 10.5, 11.2, 10.8, 9.8, 12.2, 11.0, and 9.9 cm. The average penetration depth of the comparable standard shell is 10.0 cm. We wish to know whether the new type shells penetrate farther on the average than the standard type shells.

If we compute the arithmetic mean of the ten shells, we find it is 10.70 cm. Our first impulse might be to state that on the average the new shell will penetrate 0.7 cm. farther than the standard shell. This, indeed, is our best single guess, but how sure can we be that this actually is close to the truth? One thing that might catch our notice is the variability in the individual penetration depths of the new shells. They range from 9.8 cm. to 12.2 cm. The standard deviation as measured by s calculated from the sample is 0.73 cm. Might not our sample of ten shells have contained some atypical ones of the new type which have unusually high penetrating power? Could it be that the new shell is, on the average, no better than the standard one? If we were obliged to decide, on the basis of the results obtained from these ten shells alone, whether to keep on making the standard shells or to convert our equipment to making the new shell, how can we make a valid choice?

A very worthwhile step toward a solution in such situations is to compute, from the data in hand, a confidence interval for the unknown value of the population parameter of interest. The procedure (given in Par. 2-1.4) applied to the foregoing depth-of-penetration data for the new type of shell yields the interval from 10.18 to 11.22 cm. as a 95% confidence interval for the population mean depth of penetration of shells of the new type. Inasmuch as this interval lies entirely to the right of the mean for the standard shell, 10.00 cm., we are justified in concluding that the new shell *is*, on the average, better than the standard, with only a 5% risk of being in error. Nevertheless, taking other considerations into account (e.g., cost of the new type, cost of changing over, etc.), we may conclude finally that the improvement—which may be as little as 0.18 cm., and probably not more than 1.22 cm.—is not sufficient to warrant conversion to the new type. On the other hand, the evidence that the new type is almost certainly better plus the prospect that

the improvement may be as great as 1.22 cm. may serve to recommend further developmental activity in the direction "pioneered" by the new type.

A somewhat different approach, which provides a direct answer to our question "Could it be that the new shell is on the average no better than the standard?" but not to the question of whether to convert to the new type, is to carry out a so-called *test of significance* (or test of a statistical hypothesis). In the case of the foregoing example, the formal procedure for the corresponding test of significance (Par. 3-2.2.1) turns out to be equivalent (as explained in ORDP 20-113, Chapter 21) to noting whether or not the confidence interval computed does or does not include the population mean for the standard shell (10.0 cm.). If, as in the present instance, the population mean for the standard shell is *not* included, this is taken to be a *negative* answer to our question. In other words, this is taken to be conclusive evidence (at the *5% level of significance*) *against* the *null hypothesis* that "the new shell is on the average *no better* than the standard." Rejection of the null hypothesis in this case is equivalent to accepting the indefinite *alternative hypothesis* that "the new shell *is better* on the average than the standard." If, on the other hand, the population mean for the standard shell *is* included in the confidence interval, this is taken as an *affirmative* answer to our question—not in the positive sense of definitely confirming the null hypothesis ("is no better"), but in the more-or-less neutral sense of the absence of conclusive evidence to the contrary.

As the foregoing example illustrates, an advantage of the confidence-interval approach to a decision problem is that the confidence interval gives an indication of how large the difference, if any, is likely to be, and thus provides some of the additional information usually needed to reach a final decision on the action to be taken next. For many purposes, this is a real advantage of confidence intervals over tests of significance.

However, all statistical decision problems are not amenable to solution via confidence intervals. For instance, the question at issue may be whether or not two particular characteristics of shell performance are mutually independent. In such a situation, any one of a variety of tests of significance can be used to test the null hypothesis of "no dependence." Some of these may have a reasonably good chance of rejecting the null hypothesis, and thus "discovering" the existence of a dependence when a dependence really exists—even though the exact nature of the dependence, if any, is not understood and a definitive measure of the extent of the dependence in the population is lacking.

A precise test of significance will be possible if: (a) the sampling distribution of some sample statistic is known (at least to a good approximation) for the case of "no dependence"; and (b) the effect of dependence on this statistic is known (e.g., tends to make it larger). For a confidence-interval approach to be possible, two conditions are necessary: (a) there must be agreement on what constitutes the proper measure (parameter) of dependence of the two characteristics in the population; and, (b) there must be a sample estimate of this dependence parameter whose sampling distribution is known, to a good approximation at least, for all values of the parameter. Confidence intervals tend to provide a more complete answer to statistical decision problems when they are available, but tests of significance are of wider applicability.

1-9.2 CHOICE OF NULL AND ALTERNATIVE HYPOTHESES

A statistical test always involves a *null hypothesis*, which is considered to be the hypothesis under test, as against a class of *alternative hypotheses*. The null hypothesis acts as a kind of "origin" or "base" (in the sense of "base line"), from which the alternative hypotheses deviate in one way or another to greater and lesser degrees. Thus, in the case of the classical problem of the tossing of a coin, the null or base hypothesis

specifies that the probability of "heads" on any single trial equals 1/2. If, in a particular situation, the occurrence of "heads" were an *advantage*, then we might be particularly interested in the *one-sided* class of alternative hypotheses that the probability of "heads" on any single trial equals P, where P is some (unknown) fraction exceeding 1/2. If neither "heads" nor 'tails" were intrinsically advantageous, but a bias in favor of either could be employed to advantage, then we could probably be interested in the more general *two-sided* class of alternative hypotheses specifying that the probability of "heads" on any single toss equals P, where P is some fraction (less than, or greater than, but) *not* equal to 1/2.

The important point is that the null hypothesis serves as an origin or base. In the coin-tossing instance, it also happens to be a favored, or traditional, hypothesis. This is merely a characteristic of the example selected. Indeed, the null hypothesis is often the very antithesis of what we would really like to be the case.

1-9.3 TWO KINDS OF ERRORS

In basing decisions on the outcomes of statistical tests, we always run the risks of making either one or the other of two types of error. If we reject the null hypothesis when it is true, e.g., announce a difference which really does not exist, then we make an *Error of the First Kind.* If we fail to reject a null hypothesis when it is false, e.g., fail to find an improvement in the new shell over the old when an improvement exists, then we make what is called an *Error of the Second Kind.* Although we do not know in a given instance whether we have made an error of either kind, we can know the *probability* of making either type of error.

1-9.4 SIGNIFICANCE LEVEL AND OPERATING CHARACTERISTIC (OC) CURVE OF A STATISTICAL TEST

The risk of making an error of the first kind, α, equals what is by tradition called the *level of significance* of the test. The risk of making an error of the second kind, β, varies, as one would expect, with the magnitude of the real difference, and is summarized by the *Operating Characteristic (OC) Curve* of the test. See, for example, Figure 3-5. Also, the risk β of making an error of the second kind increases as the risk α of making an error of the first kind decreases. Compare Figure 3-5 with Figure 3-6. Only with "large" samples can we "have our cake and eat it too"—and then there is the cost of the test to worry about.

1-9.5 CHOICE OF THE SIGNIFICANCE LEVEL

The significance level of a statistical test is essentially an expression of our reluctance to give up or "reject" the null hypothesis. If we adopt a "stiff" significance level, 0.01 or even 0.001, say, this implies that we are very unwilling to reject the null hypothesis unjustly. A consequence of our ultraconservatism in this respect will usually be that the probability of not rejecting the null hypothesis when it is really false will be large unless the actual deviation from the null hypothesis is large. This is clearly an entirely satisfactory state of affairs if we are *quite satisfied with the status quo* and are only interested in making a change if the change represents a very substantial improvement. For example, we may be quite satisfied with the performance of the standard type of shell in all respects, and not be willing to consider changing to the new type unless the mean depth of penetration of the new type were at least, say, 20% better (12.0 cm.).

On the other hand, the standard shell may be unsatisfactory in a number of respects and the question at issue may be whether the new type shows promise of being able to replace it, either "as is" or with further development. Here "rejection" of the null hypothesis would not imply necessary abandonment of the standard type and shifting over to the new type, but merely that the new type shows "promise" and warrants further investigation. In such a situation,

one could afford a somewhat higher risk of rejecting the null hypothesis falsely, and would take $\alpha = 0.05$ or 0.10 (or even 0.20, perhaps), in the interest of increasing the chances of detecting a small but promising improvement with a small-scale experiment. In such exploratory work, it is often more important to have a good chance of detecting a small but promising improvement than to protect oneself against crying "wolf, wolf" occasionally—because the "wolf, wolf" will be found out in due course, but a promising approach to improvement could be lost forever.

In summary, the significance level α of a statistical test should be chosen in the light of the attending circumstances, including costs. We are sometimes limited in the choice of significance level by the availability of necessary tables for some statistical tests. Two values of α, $\alpha = .05$ and $\alpha = .01$, have been most frequently used in research and development work; and are given in tabulations of test statistics. We have adopted these "standard" levels of significance for the purposes of this handbook.

1-9.6 A WORD OF CAUTION

Many persons who regularly employ statistical tests in the interpretation of research and development data do not seem to realize that all probabilities associated with such tests are calculated on the supposition that some definite set of conditions prevails. Thus, α, the level of significance (or probability of an error of the first kind), is computed on the assumption that the null hypothesis is strictly true in all respects; and β, the risk of an error of the second kind, is computed on the assumption that a particular specific alternative to the null hypothesis is true *and* that the statistical test concerned is carried out at the α-level of significance. Consequently, whatever may be the actual outcome of a statistical test, it is mathematically impossible to infer from the outcome anything whatsoever about the odds for or against some particular set of conditions being the truth.

Indeed, it is astonishing how often erroneous statements of the type "since r exceeds the 1% level of significance, the odds are 99 to 1 that there *is* a correlation between the variables" occur in research literature. How ridiculous this type of reasoning can be is brought out by the following simple example[5]: The *American Experience Mortality Table* gives .01008 as the probability of an individual aged 41 dying within the year. If we accept this table as being applicable to living persons today (which is analogous to accepting the published tables of the significance levels of tests which we apply to our data), and *if* a man's age really is 41, then the odds *are* 99 to 1 that he will live out the year. On the other hand, if we accept the table and happen to hear that some prominent individual has just died, then we *cannot* (and *would not*) conclude that the odds are 99 to 1 that his age was different from 41.

Suppose, on the other hand, that in some official capacity it is our practice to check the accuracy of age statements of all persons who say they are 41 and *then* die within the year. This practice (assuming the applicability of the American Experience Mortality Table) will lead us in the long run to suspect unjustly the word of one person in 100 whose age *was* 41, who told us so, and who then was unfortunate enough to die within the year. The *level of significance* of the test is in fact 0.01008 (1 in 100). On the other hand, this practice will also lead us to discover mis-statements of age of *all* persons professing to be 41 who are really some other age *and* who happen to die within the year. The probabilities of our discovering such mis-statements will depend on the actual ages of the persons making them. We shall, however, let slip by as correct all statements "age 41" corresponding to individuals who *are not* 41 but who do not happen to die within the year.

The moral of this is that all statistical tests can and should be viewed in terms of the consequences which may be expected to ensue from their repeated use in suitable circumstances. When viewed in this light, the great risks involved in drawing conclusions from exceedingly small samples becomes manifest to anyone who takes the time to study the OC curves for the statistical tests in common use.

REFERENCES

1. W. G. Cochran, F. Mosteller, and J. W. Tukey, "Principles of Sampling," *Journal of the American Statistical Association*, Vol. 49, pp. 13-35, 1954. (Copies of this article can be obtained from the American Statistical Association, 1757 K St., N.W., Washington 6, D. C. Price: 50 cents.)
2. W. G. Cochran, *Sampling Techniques*, John Wiley & Sons, Inc., New York, N. Y., 1953.
3. L. H. C. Tippett, *Random Sampling Numbers*, Tracts for Computers, No. 15, Cambridge University Press, 1927.
4. The Rand Corporation, *A Million Random Digits*, The Free Press, Glencoe, Ill., 1955.
5. C. Eisenhart, "The Interpretation of Tests of Significance," *Bulletin of the American Statistical Association*, Vol. 2, No. 3, pp. 79-80, April, 1941.

SOME RECOMMENDED ELEMENTARY TEXTBOOKS

A. H. Bowker and G. J. Lieberman, *Engineering Statistics*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1959.

W. J. Dixon and F. J. Massey, Jr., *Introduction to Statistical Analysis* (2d edition), McGraw-Hill Book Co., Inc., New York, N. Y., 1957.

M. J. Moroney, *Facts from Figures*, Penguin Books, Inc., Baltimore, Md., 1951.

L. H. C. Tippett, *The Methods of Statistics*, 4th edition), John Wiley & Sons, Inc., New York, N. Y., 1952.

W. A. Wallis and H. V. Roberts, *Statistics, A New Approach*, The Free Press, Glencoe, Ill., 1956.

2

STATISTICAL CONCEPTS IN METROLOGY*†

Harry H. Ku

STATISTICAL CONCEPTS OF A MEASUREMENT PROCESS

Arithmetic Numbers and Measurement Numbers

In metrological work, digital numbers are used for different purposes and consequently these numbers have different interpretations. It is therefore important to differentiate the two types of numbers which will be encountered.

*By Harry H. Ku, Statistical Engineering Laboratory, Institute for Basic Standards, National Bureau of Standards.

Chapter 2. Reprinted from the Handbook of Industrial Metrology, American Society of Tool and Manufacturing Engineers, pp. 20-50. Prentice-Hall, Inc., New York, 1967.

Arithmetic numbers are exact numbers. 3, $\sqrt{2}$, $\frac{1}{3}$, e, or π are all exact numbers by definition, although in expressing some of these numbers in digital form, approximation may have to be used. Thus, π may be written as 3.14 or 3.1416, depending on our judgment of which is the proper one to use from the combined point of view of accuracy and convenience. By the usual rules of rounding, the approximations do not differ from the exact values by more than ± 0.5 units of the last recorded digit. The accuracy of the result can always be extended if necessary.

Measurement numbers, on the other hand, are not approximations to exact numbers, but numbers obtained by operation under approximately the same conditions. For example, three measurements on the diameter of a steel shaft with a micrometer may yield the following results:

No.	*Diameter in cm*	*General notation*
1	0.396	x_1
2	0.392	x_2
3	0.401	x_3
	Sum 1.189	$\sum_{i=1}^{n} x_i$
	Average 0.3963	$\bar{x} = \frac{1}{n} \sum_{1}^{n} x_i$
	Range 0.009	$R = x_{max} - x_{min}$

There is no rounding off here. The last digit in the measured value depends on the instrument used and our ability to read it. If we had used a coarser instrument, we might have obtained 0.4, 0.4, and 0.4; if a finer instrument, we might have been able to record to the fifth digit after the decimal point. In all cases, however, the last digit given certainly does not imply that the measured value differs from the diameter D by less than ± 0.5 unit of the last digit.

Thus we see that measurement numbers differ by their very nature from arithmetic numbers. In fact, the phrase "significant figures" has little meaning in the manipulation of numbers resulting from measurements. Reflection on the simple example above will help to convince one of this fact.

Computation and Reporting of Results. By experience, the metrologist can usually select an instrument to give him results adequate for his needs, as illustrated in the example above. Unfortunately, in the process of computation, both arithmetic numbers and measurement numbers are present, and frequently confusion reigns over the number of digits to be kept in successive arithmetic operations.

No general rule can be given for all types of arithmetic operations. If the instrument is well-chosen, severe rounding would result in loss of information. One suggestion, therefore, is to treat all measurement numbers as exact numbers in the operations and to round off the final result only.

Another recommended procedure is to carry two or three extra figures throughout the computation, and then to round off the final reported value to an appropriate number of digits.

The "appropriate" number of digits to be retained in the final result depends on the "uncertainties" attached to this reported value. The term "uncertainty" will be treated later under "Precision and Accuracy"; our only concern here is the number of digits in the expression for uncertainty.

A recommended rule is that the uncertainty should be stated to no more than two significant figures, and the reported value itself should be stated to the last place affected by the qualification given by the uncertainty statement. An example is:

> "The apparent mass correction for the nominal 10 g weight is +0.0420 mg with an overall uncertainty of ±0.0087 mg using three standard deviations as a limit to the effect of random errors of measurement, the magnitude of systematic errors from known sources being negligible."

The sentence form is preferred since then the burden is on the reporter to specify exactly the meaning of the term uncertainty, and to spell out its components. Abbreviated forms such as $a \pm b$, where a is the reported value and b a measure of uncertainty in some vague sense, should always be avoided.

Properties of Measurement Numbers

The study of the properties of measurement numbers, or the Theory of Errors, formally began with Thomas Simpson more than two hundred years ago, and attained its full development in the hands of Laplace and Gauss. In the next subsections some of the important properties of measurement numbers will be discussed and summarized, thus providing a basis for the statistical treatment and analysis of these numbers in the following major section.

The Limiting Mean. As shown in the micrometer example above, the results of *repeated measurements of a single physical quantity under essentially the same conditions* yield a set of measurement numbers. Each member of this set is an estimate of the quantity being measured, and has equal claims on its value. By convention, the numerical values of these n measurements are denoted by $x_1, x_2, \ldots, x_n$, the arithmetic mean by $\bar{x}$, and the range by R, i.e., the difference between the largest value and the smallest value obtained in the n measurements.

If the results of measurements are to make any sense for the purpose at hand, we must require these numbers, though different, to behave as a group in a certain predictable manner. Experience has shown that this is

indeed the case under the conditions stated in italics above. In fact, let us adopt as the postulate of measurement a statement due to N. Ernest Dorsey (reference 2)*:

> "The mean of a family of measurements—of a number of measurements for a given quantity carried out by the same apparatus, procedure, and observer—approaches a definite value as the number of measurements is indefinitely increased. Otherwise, they could not properly be called measurements of a given quantity. In the theory of errors, this limiting mean is frequently called the 'true' value, although it bears no necessary relation to the true quaesitum, to the actual value of the quantity that the observer desires to measure. This has often confused the unwary. Let us call it the limiting mean."

Thus, according to this postulate, there exists a limiting mean m to which $\bar{x}$ approaches as the number of measurements increases indefinitely, or, in symbols $\bar{x} \rightarrow m$ as $n \rightarrow \infty$. Furthermore, if the true value is τ, there is usually a difference between m and τ, or $\Delta = m - \tau$, where Δ is defined as the bias or systematic error of the measurements.

In practice, however, we will run into difficulties. The value of m cannot be obtained since one cannot make an infinite number of measurements. Even for a large number of measurements, the conditions will not remain constant, since changes occur from hour to hour, and from day to day. The value of τ is unknown and usually unknowable, hence also the bias. Nevertheless, this seemingly simple postulate does provide a sound foundation to build on toward a mathematical model, from which estimates can be made and inference drawn, as will be seen later on.

Range, Variance, and Standard Deviation. The range of n measurements, on the other hand, does not enjoy this desirable property of the arithmetic mean. With one more measurement, the range may increase but cannot decrease. Since only the largest and the smallest numbers enter into its calculation, obviously the additional information provided by the measurements in between is lost. It will be desirable to look for another measure of the dispersion (spread, or scattering) of our measurements which will utilize each measurement made with equal weight, and which will approach a definite number as the number of measurements is indefinitely increased.

A number of such measures can be constructed; the most frequently used are the variance and the standard deviation. The choice of the variance as the measure of dispersion is based upon its mathematical convenience and maneuverability Variance is defined as the value approached by the average of the sum of squares of the deviations of individual measurements from the limiting mean as the number of measurements is indefinitely

*References are listed at the end of this chapter.

increased, or in symbols:

$$\frac{1}{n} \sum (x_i - m)^2 \rightarrow \sigma^2 = \text{variance, as } n \rightarrow \infty$$

The positive square root of the variance, σ, is called the standard deviation (of a single measurement); the standard deviation is of the same dimensionality as the limiting mean.

There are other measures of dispersion, such as average deviation and probable error. The relationships between these measures and the standard deviation can be found in reference 1.

Population and the Frequency Curve. We shall call the limiting mean m the location parameter and the standard deviation σ the scale parameter of the population of measurement numbers generated by a particular measurement process. By population is meant the conceptually infinite number of measurements that can be generated. The two numbers m and σ describe this population of measurements to a large extent, and specify it completely in one important special case.

Our model of a measurement process consists then of a defined population of measurement numbers with a limiting mean m and a standard deviation σ. The result of a single measurement X^* can take randomly any of the values belonging to this population. The probability that a particular measurement yields a value of X which is less than or equal to x' is the proportion of the population that is less than or equal to x', in symbols

$$P\{X \leq x'\} = \text{proportion of population less than or equal to } x'$$

Similar statements can be made for the probability that X will be greater than or equal to x'', or for X between x' and x'' as follows: $P\{X \geq x''\}$, or $P\{x' \leq X \leq x''\}$.

For a measurement process that yields numbers on a continuous scale, the distribution of values of X for the population can be represented by a smooth curve, for example, curve C in Fig. 2-1. C is called a frequency curve. The area between C and the abscissa bounded by any two values (x_1 and x_2) is the proportion of the population that takes values between the two values, or the probability that X will assume values between x_1 and x_2. For example, the probability that $X \leq x'$, can be represented by the shaded area to the left of x'; the total area between the frequency curve and the abscissa being one by definition.

Note that the shape of C is not determined by m and σ alone. Any curve C' enclosing an area of unity with the abscissa defines the distribution of a particular population. Two examples, the uniform distribution and

*Convention is followed in using the capital X to represent the value that might be produced by employing the measurement process to obtain a measurement (i.e., a random variable), and the lower case x to represent a particular value of X observed.

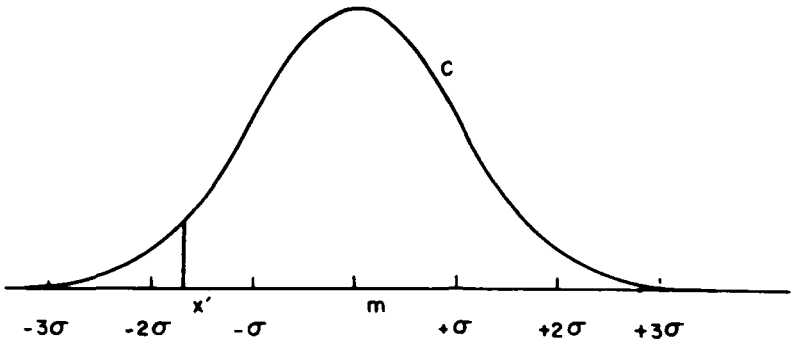
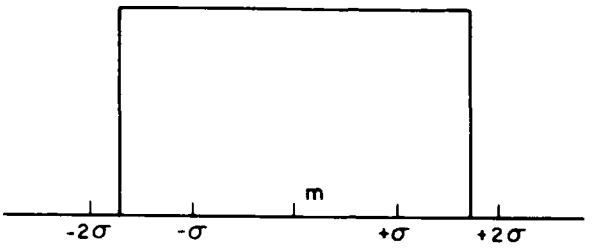


Fig. 2-1. A symmetrical distribution.



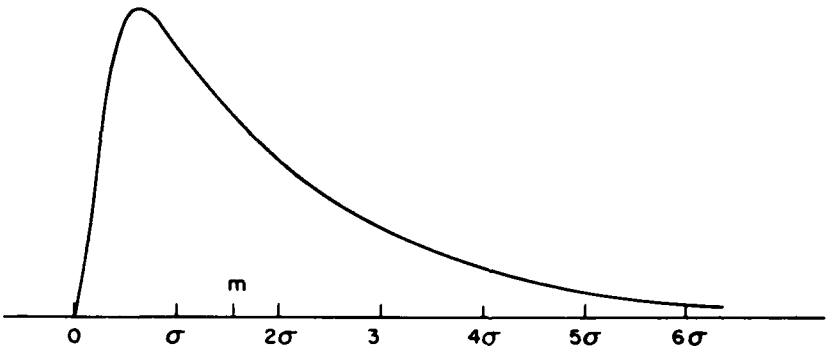


Fig. 2-2. (A) The uniform distribution. (B) The log-normal distribution.

the log-normal distribution are given in Figs. 2-2A and 2-2B. These and other distributions are useful in describing certain populations.

The Normal Distribution. For data generated by a measurement process, the following properties are usually observed:

1. The results spread roughly symmetrically about a central value.
2. Small deviations from this central value are more frequently found than large deviations.

A measurement process having these two properties would generate a frequency curve similar to that shown in Fig. 2-1 which is symmetrical and bunched together about m. The study of a particular theoretical representation of a frequency curve of this type leads to the celebrated bell-shaped normal curve (Gauss error curve.). Measurements having such a normal frequency curve are said to be normally distributed, or distributed in accordance with the normal law of error.

The normal curve can be represented exactly by the mathematical expression

$$y = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-1\,2[(x-m)^2\cdot\sigma^2]} \tag{2-0}$$

where y is the ordinate and x the abscissa and $e \doteq 2.71828$ is the base of natural logarithms.

Some of the important features of the normal curve are:

1. It is symmetrical about m.
2. The area under the curve is one, as required.
3. If σ is used as unit on the abscissa, then the area under the curve between constant multiples of σ can be computed from tabulated values of the normal distribution. In particular, areas under the curve for some useful intervals between $m - k\sigma$ and $m + k\sigma$ are given in Table 2-1. Thus about two-thirds of the area lies within one σ of m, more than 95 percent within 2σ of m, and less than 0.3 percent beyond 3σ from m.

TABLE 2-1

	Area under normal curve between $m - k\sigma$ and $m + k\sigma$					
k:	0.6745	1.00	1.96	2.00	2.58	3.00
Percent area under curve (approx.):	50.0	68.3	95.0	95.5	99.0	99.7

4. From Eq. (2-0), it is evident that the frequency curve is completely determined by the two parameters m and σ.

The normal distribution has been studied intensively during the past century. Consequently, if the measurements follow a normal distribution, we can say a great deal about the measurement process. The question remains: How do we know that this is so from the limited number of repeated measurements on hand?

The answer is that we don't! However, in most instances the metrologist may be willing

1. to assume that the measurement process generates numbers that follow a normal distribution approximately, and act as if this were so,
2. to rely on the so-called Central Limit Theorem, one version of which

is the following*: "If a population has a finite variance σ^2 and mean m, then the distribution of the sample mean (of n independent measurements) approaches the normal distribution with variance σ^2/n and mean m as the sample size n increases." This remarkable and powerful theorem is indeed tailored for measurement processes. First, every measurement process must by definition have a finite mean and variance. Second, the sample mean $\bar{x}$ is the quantity of interest which, according to the theorem, will be approximately normally distributed for large sample sizes. Third, the measure of dispersion, i.e., the standard deviation of the sample mean, is reduced by a factor of $1/\sqrt{n}$! This last statement is true in general for all measurement processes in which the measurements are "independent" and for all n. It is therefore not a consequence of the Central Limit Theorem. The theorem guarantees, however, that the distribution of sample means of *independent* measurements will be *approximately* normal with the specified limiting mean and standard deviation $\sigma/\sqrt{n}$ *for large* n.

In fact, for a measurement process with a frequency curve that is symmetrical about the mean, and with small deviations from the mean as compared to the magnitude of the quantity measured, the normal approximation to the distribution of $\bar{x}$ becomes very good even for n as small as 3 or 4. Figure 2-3 shows the uniform and normal distribution having the same mean and standard deviation. The peaked curve is actually two curves, representing the distribution of arithmetic means of four independent measurements from the respective distributions. These curves are indistinguishable to this scale.

A formal definition of the concept of "independence" is out of the scope here. Intuitively, we may say that n normally distributed measurements are independent if these measurements are not correlated or associated in any way. Thus, a sequence of measurements showing a trend or pattern are not independent measurements.

There are many ways by which dependence or correlation creeps into a set of measurement data; several of the common causes are the following:

1. Measurements are correlated through a factor that has not been considered, or has been considered to be of no appreciable effect on the results.
2. A standard correction constant has been used for a factor, e.g., temperature, but the constant may overcorrect or undercorrect for particular samples.
3. Measurements are correlated through time of the day, between days, weeks, or seasons.

*From Chapter 7, *Introduction to the Theory of Statistics*, by A. M. Mood, McGraw-Hill Book Company, New York, 1950.

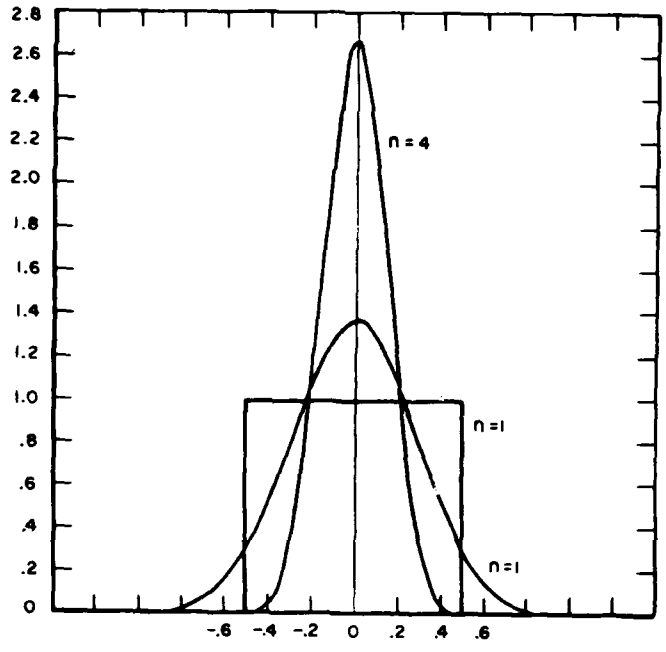


Fig. 2-3. Uniform and normal distribution of individual measurements having the same mean and standard deviation, and the corresponding distribution(s) of arithmetic means of four independent measurements.

4. Measurements are correlated through rejection of valid data, when the rejection is based on the size of the number in relation to others of the group.

The traditional way of plotting the data in the sequence they are taken, or in some rational grouping, is perhaps still the most effective way of detecting trends or correlation.

Estimates of Population Characteristics. In the above section it is shown that the limiting mean m and the variance σ^2 completely specify a measurement process that follows the normal distribution. In practice, m and σ^2 are not known and cannot be computed from a finite number of measurements. This leads to the use of the sample mean $\bar{x}$ as an estimate of the limiting mean m and s^2, the square of the computed standard deviation of the sample, as an estimate of the variance. The standard deviation of the average of n measurements, $\sigma/\sqrt{n}$, is sometimes referred to as the standard error of the mean, and is estimated by $s/\sqrt{n}$.

We note that the making of n independent measurements is equivalent

to drawing a sample of size n at random from the population of measurements. Two concepts are of importance here:

1. The measurement process is established and under control, meaning that the limiting mean and the standard deviation do possess definite values which will not change over a reasonable period of time.
2. The measurements are randomly drawn from this population, implying that the values are of equal weights, and there is no prejudice in the method of selection. Suppose out of three measurements the one which is far apart from the other two is rejected, then the result will not be a random sample.

For a random sample we can say that $\bar{x}$ is an unbiased estimate of m, and s^2 is an unbiased estimate of σ^2, i.e., the limiting mean of $\bar{x}$ is equal to m and of s^2 to σ^2, where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right]$$

In addition, we define

$$s = \sqrt{s^2} = \text{computed standard deviation}$$

Examples of numerical calculations of $\bar{x}$ and s^2 and s are shown in Tables 2-5 and 2-6.

Interpretation and Computation of Confidence Interval and Limits

By making k sets of n measurements each, we can compute and arrange k, $\bar{x}$'s, and s's in a tabular form as follows:

Set	*Sample mean*	*Sample standard deviation*
1	$\bar{x}_1$	s_1
2	$\bar{x}_2$	s_2
.	.	.
.	.	.
.	.	.
j	$\bar{x}_j$	s_j
.	.	.
.	.	.
.	.	.
k	$\bar{x}_k$	s_k

In the array of $\bar{x}$'s, no two will be likely to have exactly the same value. From the Central Limit Theorem it can be deduced that the $\bar{x}$'s will be

approximately normally distributed with standard deviation $\sigma/\sqrt{n}$. The frequency curve of $\bar{x}$ will be centered about the limiting mean m and will have the scale factor $\sigma/\sqrt{n}$. In other words, $\bar{x} - m$ will be centered about zero, and the quantity

$$z = \frac{\bar{x} - m}{\sigma/\sqrt{n}}$$

has the properties of a single observation from the "standardized" normal distribution which has a mean of zero and a standard deviation of one.

From tabulated values of the standardized normal distribution it is known that 95 percent of z values will be bounded between -1.96 and $+1.96$. Hence the statement

$$-1.96 < \frac{\bar{x} - m}{\sigma/\sqrt{n}} < +1.96$$

or its equivalent,

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < m < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

will be correct 95 percent of the time in the long run. The interval $\bar{x} - 1.96(\sigma/\sqrt{n})$ to $\bar{x} + 1.96(\sigma/\sqrt{n})$ is called a *confidence interval* for m. The probability that the confidence interval will cover the limiting mean, 0.95 in this case, is called the confidence level or confidence coefficient. The values of the end points of a confidence interval are called confidence limits. It is to be borne in mind that $\bar{x}$ will fluctuate from set to set, and the interval calculated for a particular $\bar{x}_j$ may or may not cover m.

In the above discussion we have selected a two-sided interval symmetrical about $\bar{x}$. For such intervals the confidence coefficient is usually denoted by $1 - \alpha$, where $\alpha/2$ is the percent of the area under the frequency curve of z that is cut off from each tail.

In most cases, σ is not known and an estimate of σ is computed from the same set of measurements we use to calculate $\bar{x}$. Nevertheless, let us form a quantity similar to z, which is

$$t = \frac{\bar{x} - m}{s/\sqrt{n}}$$

and if we know the distribution of t, we could make the same type of statement as before. In fact the distribution of t is known for the case of normally distributed measurements.

The distribution of t was obtained mathematically by William S. Gosset under the pen name of "Student," hence the distribution of t is called the Student's distribution. In the expression for t, both $\bar{x}$ and s fluctuate from set to set of measurements. Intuitively we will expect the value of t to be larger than that of z for a statement with the same probability of being correct. This is indeed the case. The values of t are listed in Table 2-2.

TABLE 2-2. A BRIEF TABLE OF VALUES OF t*

Degrees of freedom	Confidence Level: $1-\alpha$			
ν	0.500	0.900	0.950	0.990
1	1.000	6.314	12.706	63.657
2	.816	2.920	4.303	9.925
3	.765	2.353	3.182	5.841
4	.741	2.132	2.776	4.604
5	.727	2.015	2.571	4.032
6	.718	1.943	2.447	3.707
7	.711	1.895	2.365	3.499
10	.700	1.812	2.228	3.169
15	.691	1.753	2.131	2.947
20	.687	1.725	2.086	2.845
30	.683	1.697	2.042	2.750
60	.679	1.671	2.000	2.660
∞	.674	1.645	1.960	2.576

*Adapted from *Biometrika Tables for Statisticians*, Vol. I, edited by E. S. Pearson and H. O. Hartley, The University Press, Cambridge, 1958.

To find a value for t, we need to know the "degrees of freedom" (ν) associated with the computed standard deviation s. Since $\bar{x}$ is calculated from the same n numbers and has a fixed value, the nth value of x_i is completely determined by $\bar{x}$ and the other $(n-1)x$ values. Hence the degrees of freedom here are $n-1$.

Having the table for the distribution of t, and using the same reasoning as before, we can make the statement that

$$\bar{x} - t\frac{s}{\sqrt{n}} < m < \bar{x} + t\frac{s}{\sqrt{n}}$$

and our statement will be correct $100(1-\alpha)$ percent of the time in the long run. The value of t depends on the degrees of freedom ν and the probability level. From the table, we get for a confidence level of 0.95, the following lower and upper confidence limits:

ν	$L_l = \bar{x} - t(s/\sqrt{n})$	$L_u = \bar{x} + t(s/\sqrt{n})$
1	$\bar{x} - 12.706(s/\sqrt{n})$	$\bar{x} + 12.706(s/\sqrt{n})$
2	$\bar{x} - 4.303(s/\sqrt{n})$	$\bar{x} + 4.303(s/\sqrt{n})$
3	$\bar{x} - 3.182(s/\sqrt{n})$	$\bar{x} + 3.182(s/\sqrt{n})$

The value of t for $\nu = \infty$ is 1.96, the same as for the case of known σ. Notice that very little can be said about m with two measurements. However, for n larger than 2, the interval predicted to contain m narrows down steadily, due to both the smaller value of t and the divisor $\sqrt{n}$.

It is probably worthwhile to emphasize again that each particular confidence interval computed as a result of n measurements will either include m or fail to include m. The probability statement refers to the fact that if we make a long series of sets of n measurements, and if we compute a confidence interval for m from each set by the prescribed method, we would expect 95 percent of such intervals to include m.

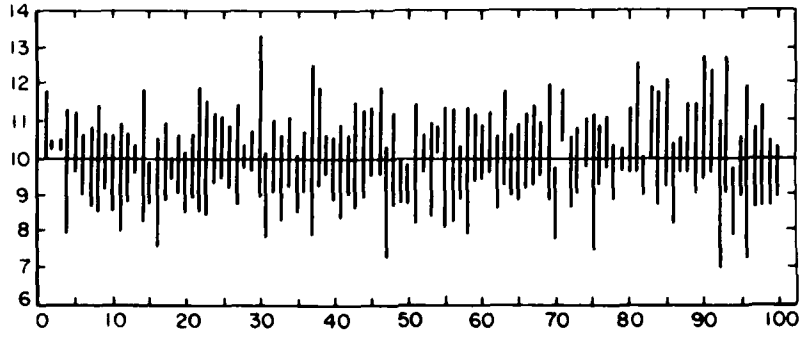


Fig. 2-4. Computed 90% confidence intervals for 100 samples of size 4 drawn at random from a normal population with $m = 10$, $\sigma = 1$.

Figure 2-4 shows the 90 percent confidence intervals ($P = 0.90$) computed from 100 samples of $n = 4$ from a normal population with $m = 10$, and $\sigma = 1$. Three interesting features are to be noted:

1. The number of intervals that include m actually turns out to be 90, the expected number.
2. The surprising variation of the sizes of these intervals.
3. The closeness of the mid-points of these intervals to the line for the mean does not seem to be related to the spread. In samples No. 2 and No. 3, the four values must have been very close together, but both of these intervals failed to include the line for the mean.

From the widths of computed confidence intervals, one may get an intuitive feeling whether the number of measurements n is reasonable and sufficient for the purpose on hand. It is true that, even for small n, the confidence intervals will cover the limiting mean with the specified probability, yet the limits may be so far apart as to be of no practical significance. For detecting a specified magnitude of interest, e.g., the difference between two means, the approximate number of measurements required can be solved by equating the half-width of the confidence interval to this difference and solving for n, using σ when known, or using s by trial and error if σ is not known. Tables of sample sizes required for certain prescribed conditions are given in reference 4.

Index of Precision. Since σ is a measure of the spread of the frequency curve about the limiting mean, σ may be defined as an index of precision. Thus a measurement process with a standard deviation σ_1 is said to be more precise than another with a standard deviation σ_2 if σ_1 is smaller than σ_2. (In fact, σ is really a measure of imprecision since the imprecision is directly proportional to σ.)

Consider the means of sets of n independent measurements as a new derived measurement process. The standard deviation of the new process is $\sigma/\sqrt{n}$. It is therefore possible to derive from a less precise measurement process a new process which has a standard deviation equal to that of a more precise process. This is accomplished by making more measurements.

Suppose $m_1 = m_2$, but $\sigma_1 = 2\sigma_2$. Then for a derived process to have $\sigma_1' = \sigma_2$, we need

$$\sigma_1' = \frac{\sigma_1}{\sqrt{n}} = \frac{2\sigma_2}{\sqrt{4}}$$

or we need to use the average of four measurements as a single measurement. Thus for a required degree of precision, the number of measurements, n_1 and n_2, needed for measurement processes I and II is proportional to the squares of their respective standard deviations (variances), or in symbols

$$\frac{n_1}{n_2} = \frac{\sigma_1^2}{\sigma_2^2}$$

If σ is not known, and the best estimate we have of σ is a computed standard deviation s based on n measurements, then s could be used as an estimate of the index of precision. The value of s, however, may vary considerably from sample to sample in the case of a small number of measurements as was shown in Fig. 2-4, where the lengths of the intervals are constant multiples of s computed from the samples. The number n or the degrees of freedom ν must be considered along with s in indicating how reliable an estimate s is of σ. In what follows, whenever the terms standard deviation about the limiting mean (σ), or standard error of the mean ($\sigma_{\bar{x}}$), are used, the respective estimates s and $s/\sqrt{n}$ may be substituted, by taking into consideration the above reservation.

In metrology or calibration work, the precision of the reported value is an integral part of the result. In fact, precision is the main criterion by which the quality of the work is judged. Hence, the laboratory reporting the value must be prepared to give evidence of the precision claimed. Obviously an estimate of the standard deviation of the measurement process based only on a small number of measurements cannot be considered as convincing evidence. By the use of the control chart method for standard deviation and by the calibration of one's own standard at frequent intervals, as

subsequently described, the laboratory may eventually claim that the standard deviation is in fact known and the measurement process is stable, with readily available evidence to support these claims.

Interpretation of Precision. Since a measurement process generates numbers as the results of repeated measurements of a single physical quantity under essentially the same conditions, the method and procedure in obtaining these numbers must be specified in detail. However, no amount of detail would cover all the contingencies that may arise, or cover all the factors that may affect the results of measurement. Thus a single operator in a single day with a single instrument may generate a process with a precision index measured by σ. Many operators measuring the same quantity over a period of time with a number of instruments will yield a precision index measured by σ'. Logically σ' must be larger than σ, and in practice it is usually considerably larger. Consequently, modifiers of the words "precision" are recommended by ASTM* to qualify in an unambiguous manner what is meant. Examples are "single-operator-machine," "multi-laboratory," "single-operator-day," etc. The same publication warns against the use of the terms "repeatability" and "reproducibility" if the interpretation of these terms is not clear from the context.

The standard deviation σ or the standard error $\sigma/\sqrt{n}$ can be considered as a yardstick with which we can gage the difference between two results obtained as measurements of the same physical quantity. If our interest is to compare the results of one operator against another, the single-operator precision is probably appropriate, and if the two results differ by an amount considered to be large as measured by the standard errors, we may conclude that the evidence is predominantly against the two results being truly equal. In comparing the results of two laboratories, the single-operator precision is obviously an inadequate measure to use, since the precision of each laboratory must include factors such as multi-operator-day-instruments.

Hence the selection of an index of precision depends strongly on the purposes for which the results are to be used or might be used. It is common experience that three measurements made within the hour are closer together than three measurements made on, say, three separate days. However, an index of precision based on the former is generally not a justifiable indicator of the quality of the reported value. For a thorough discussion on the *realistic* evaluation of precision see Section 4 of reference 2.

Accuracy. The term "accuracy" usually denotes in some sense the closeness of the measured values to the true value, taking into consideration

*"Use of the Terms Precision and Accuracy as Applied to the Measurement of a Property of a Material," ASTM Designation, E177-61T, 1961.

both precision and bias. Bias, defined as the difference between the limiting mean and the true value, is a constant, and does not behave in the same way as the index of precision, the standard deviation. In many instances, the possible sources of biases are known but their magnitudes and directions are not known. The overall bias is of necessity reported in terms of estimated bounds that reasonably include the combined effect of all the elemental biases. Since there are no accepted ways to estimate bounds for elemental biases, or to combine them, these should be reported and discussed in sufficient detail to enable others to use their own judgment on the matter.

It is recommended that an index of accuracy be expressed as a pair of numbers, one the credible bounds for bias, and the other an index of precision, usually in the form of a multiple of the standard deviation (or estimated standard deviation). The terms "uncertainty" and "limits of error" are sometimes used to express the sum of these two components, and their meanings are ambiguous unless the components are spelled out in detail.

STATISTICAL ANALYSIS OF MEASUREMENT DATA

In the last section the basic concepts of a measurement process were given in an expository manner. These concepts, necessary to the statistical analysis to be presented in this section, are summarized and reviewed below. By making a measurement we obtain a number intended to express quantitatively a measure of "the property of a thing." Measurement numbers differ from ordinary arithmetic numbers, and the usual "significant figure" treatment is not appropriate. Repeated measurement of a single physical quantity under essentially the same conditions generates a sequence of numbers $x_1, x_2, \ldots, x_n$. A measurement process is established if this conceptually infinite sequence has a limiting mean m and a standard deviation σ.

For many measurement processes encountered in metrology, the sequence of numbers generated follows approximately the normal distribution, specified completely by the two quantities m and σ. Moreover, averages of n independent measurement numbers tend to be normally distributed with the limiting mean m and the standard deviation $\sigma/\sqrt{n}$, regardless of the distribution of the original numbers. Normally distributed measurements are independent if they are not correlated or associated in any way. A sequence of measurements showing a trend or pattern are not independent measurements. Since m and σ are usually not known, these quantities are estimated by calculating $\bar{x}$ and s from n measurements, where

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i$$

and

$$s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\left[\sum_{1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}\right]}$$

The distribution of the quantity $t = (\bar{x} - m)/(s/\sqrt{n})$ (for x normally distributed) is known. From the tabulated values of t (see Table 2-2), confidence intervals can be constructed to bracket m for a given confidence coefficient $1 - \alpha$ (probability of being correct in the long run).

The confidence limits are the end points of confidence intervals defined by

$$L_l = \bar{x} - t\frac{s}{\sqrt{n}}$$

$$L_u = \bar{x} + t\frac{s}{\sqrt{n}}$$

where the value of t is determined by two parameters, namely, the degrees of freedom ν associated with s and the confidence coefficient $1 - \alpha$.

The width of a confidence interval gives an intuitive measure of the uncertainty of the evidence given by the data. Too wide an interval may merely indicate that more measurements need to be made for the objective desired.

Algebra for the Manipulation of Limiting Means and Variances

Basic Formulas. A number of basic formulas are extremely useful in dealing with a quantity which is a combination of other measured quantities.

1. Let m_x and m_y be the respective limiting means of two measured quantities X and Y, and a, b be constants, then

$$\left.\begin{aligned} m_{x+y} &= m_x + m_y \\ m_{x-y} &= m_x - m_y \\ m_{ax+by} &= am_x + bm_y \end{aligned}\right\} \quad (2\text{-}1)$$

2. If, in addition, X and Y are independent, then it is also true that

$$m_{xy} = m_x m_y \quad (2\text{-}2)$$

For paired values of X and Y, we can form the quantity Z, with

$$Z = (X - m_x)(Y - m_y) \quad (2\text{-}3)$$

Then by formula (2-2) for independent variables,

$$\begin{aligned} m_z &= m_{(x-m_x)}m_{(y-m_y)} \\ &= (m_x - m_x)(m_y - m_y) = 0 \end{aligned}$$

Thus $m_z = 0$ when X and Y are independent.

3. The limiting mean of Z in (2-3) is defined as the covariance of X and Y and is usually denoted by cov (X, Y), or σ_{xy}. The covariance, similar to the variance, is estimated by

$$s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) \tag{2-4}$$

Thus if X and Y are correlated in such a way that paired values are likely to be both higher or lower than their respective means, then s_{xy} tends to be positive. If a high x value is likely to be paired with a low y value, and vice versa, then s_{xy} tends to be negative. If X and Y are not correlated, s_{xy} tends to zero (for large n).

4. The correlation coefficient ρ is defined as:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{2-5}$$

and is estimated by

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{2-6}$$

Both ρ and r lie between -1 and $+1$.

5. Let σ_x^2 and σ_y^2 be the respective variances of X and Y, and σ_{xy} the covariance of X and Y, then

$$\begin{aligned} \sigma_{x+y}^2 &= \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy} \\ \sigma_{x-y}^2 &= \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy} \end{aligned} \tag{2-7}$$

If X and Y are independent, $\sigma_{xy} = 0$, then

$$\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 = \sigma_{x-y}^2 \tag{2-8}$$

Since the variance of a constant is zero, we have

$$\begin{aligned} \sigma_{ax+b}^2 &= a^2\sigma_x^2 \\ \sigma_{ax+by}^2 &= a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_{xy} \end{aligned} \tag{2-9}$$

In particular, if X and Y are independent and normally distributed, then $aX + bY$ is normally distributed with limiting mean $am_x + bm_y$ and variance $a^2\sigma_x^2 + b^2\sigma_y^2$.

For measurement situations in general, metrologists usually strive to get measurements that are independent, or can be assumed to be independent. The case when two quantities are dependent because both are functions of other measured quantities will be treated under propagation of error formulas (see Eq. 2-13).

6. Standard errors of the sample mean and the weighted means (of independent measurements) are special cases of the above. Since $\bar{x} = (1/n) \sum x_i$ and the x_i's are independent with variance σ_x^2, it follows, by (2-9), that

$$\sigma_{\bar{x}}^2 = \left(\frac{1}{n}\right)^2 \sigma_{x_1}^2 + \left(\frac{1}{n}\right)^2 \sigma_{x_2}^2 + \cdots \left(\frac{1}{n}\right)^2 \sigma_{x_n}^2 = \frac{\sigma_x^2}{n} \tag{2-10}$$

as previously stated.

If $\bar{x}_1$ is an average of k values, and $\bar{x}_2$ is an average of n values, then for the over-all average, $\bar{\bar{x}}$, it is logical to compute

$$\bar{\bar{x}} = \frac{x_1 + \cdots + x_k + x_{k+1} + \cdots + x_{k+n}}{k + n}$$

and $\sigma_{\bar{\bar{x}}}^2 = \sigma_x^2/(k + n)$. However, this is equivalent to a weighted mean of $\bar{x}_1$ and $\bar{x}_2$, where the weights are proportional to the number of measurements in each average, i.e.,

$$w_1 = k, \qquad w_2 = n$$

and

$$\bar{\bar{x}} = \left(\frac{w_1}{w_1 + w_2}\right)\bar{x}_1 + \left(\frac{w_2}{w_1 + w_2}\right)\bar{x}_2$$

$$= \frac{k}{n + k}\,\bar{x}_1 + \frac{n}{n + k}\,\bar{x}_2$$

Since

$$\frac{\sigma_{\bar{x}_1}^2}{\sigma_{\bar{x}_2}^2} = \frac{\sigma^2/k}{\sigma^2/n} = \frac{n}{k} = \frac{w_2}{w_1}$$

the weighting factors w_1 and w_2 are therefore also inversely proportional to the respective variances of the averages. This principle can be extended to more than two variables in the following manner.

Let $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ be a set of averages estimating the same quantity. The over-all average may be computed to be

$$\bar{\bar{x}} = \frac{1}{w_1 + w_2 + \cdots + w_k}(w_1\bar{x}_1 + w_2\bar{x}_2 + \cdots + w_k\bar{x}_k)$$

where

$$w_1 = \frac{1}{\sigma_{\bar{x}_1}^2}, \qquad w_2 = \frac{1}{\sigma_{\bar{x}_2}^2}, \qquad \ldots, \qquad w_k = \frac{1}{\sigma_{\bar{x}_k}^2}$$

The variance of $\bar{\bar{x}}$ is, by (2-9),

$$\sigma_{\bar{\bar{x}}}^2 = \frac{1}{w_1 + w_2 + \cdots + w_k} \tag{2-11}$$

In practice, the estimated variances $s_{\bar{x}}^2$ will have to be used in the above formulas, and consequently the equations hold only as approximations.

Propagation of error formulas. The results of a measurement process can usually be expressed by a number of averages $\bar{x}, \bar{y}, \ldots$, and the standard errors of these averages $s_{\bar{x}} = s_x/\sqrt{n}$, $s_{\bar{y}} = s_y/\sqrt{k}$, etc. These results, however, may not be of direct interest; the quantity of interest is in the functional relationship $m_w = f(m_x, m_y)$. It is desired to estimate m_w by $\bar{w} = f(\bar{x}, \bar{y})$ and to compute $s_{\bar{w}}$ as an estimate of $\sigma_{\bar{w}}$.

If the errors of measurements of these quantities are small in comparison with the values measured, the propagation of error formulas usually work surprisingly well. The $\sigma^2_{\bar{w}}$, $\sigma^2_{\bar{x}}$, and $\sigma^2_{\bar{y}}$ that are used in the following formulas will often be replaced in practice by the computed values $s^2_{\bar{w}}$, $s^2_{\bar{x}}$, and $s^2_{\bar{y}}$.

The general formula for $\sigma^2_{\bar{w}}$ is given by

$$\sigma^2_{\bar{w}} \doteq \left[\frac{\partial f}{\partial x}\right]^2 \sigma^2_{\bar{x}} + \left[\frac{\partial f}{\partial y}\right]^2 \sigma^2_{\bar{y}} + 2\left[\frac{\partial f}{\partial x}\right]\left[\frac{\partial f}{\partial y}\right]\rho_{\bar{x}\bar{y}}\sigma_{\bar{x}}\sigma_{\bar{y}} \qquad (2\text{-}12)$$

where the partial derivatives in square brackets are to be evaluated at the averages of x and y. If X and Y are independent, $\rho = 0$ and therefore the last term equals zero. If X and Y are measured in pairs, $s_{\bar{x}\bar{y}}$ (Eq. 2-4) can be used as an estimate of $\rho_{\bar{x}\bar{y}}\sigma_{\bar{x}}\sigma_{\bar{y}}$.

If W is functionally related to U and V by

$$m_w = f(m_u, m_v)$$

TABLE 2-3. PROPAGATION OF ERROR FORMULAS FOR SOME SIMPLE FUNCTIONS

(X and Y are assumed to be independent.)

Function form	Approximate formula for $s^2_{\bar{w}}$
$m_w = Am_x + Bm_y$	$A^2 s^2_{\bar{x}} + B^2 s^2_{\bar{y}}$
$m_w = \frac{m_x}{m_y}$	$\left(\frac{\bar{x}}{\bar{y}}\right)^2\left(\frac{s^2_{\bar{x}}}{\bar{x}^2} + \frac{s^2_{\bar{y}}}{\bar{y}^2}\right)$
$m_w = \frac{1}{m_y}$	$\frac{s^2_{\bar{y}}}{\bar{y}^4}$
$m_w = \frac{m_x}{m_x + m_y}$	$\left(\frac{\bar{w}}{\bar{x}}\right)^4(\bar{y}^2 s^2_{\bar{x}} + \bar{x}^2 s^2_{\bar{y}})$
$m_w = \frac{m_x}{1 + m_x}$	$\frac{s^2_{\bar{x}}}{(1 + \bar{x})^4}$
$*m_w = m_x m_y$	$(\bar{x}\bar{y})^2\left(\frac{s^2_{\bar{x}}}{\bar{x}^2} + \frac{s^2_{\bar{y}}}{\bar{y}^2}\right)$
$*m_w = m_x^2$	$4\bar{x}^2 s^2_{\bar{x}}$
$m_w = \sqrt{m_x}$	$\frac{1}{4}\frac{s^2_{\bar{x}}}{\bar{x}}$
$*m_w = \ln m_x$	$\frac{s^2_{\bar{x}}}{\bar{x}^2}$
$*m_w = km_x^a m_y^b$	$\bar{w}^2\left(a^2\frac{s^2_{\bar{x}}}{\bar{x}^2} + b^2\frac{s^2_{\bar{y}}}{\bar{y}^2}\right)$
$*m_w = e^{m_x}$	$e^{2\bar{x}} s^2_{\bar{x}}$
$W = 100\frac{s_{\bar{x}}}{\bar{x}}$ (=coefficient of variation)	$\frac{\bar{w}^2}{2(n-1)}$ (not directly derived from the formulas)†

*Distribution of $\bar{w}$ is highly skewed and normal approximation could be seriously in error for small n.

†See, for example, *Statistical Theory with Engineering Applications*, by A. Hald, John Wiley & Sons, Inc., New York, 1952, p. 301.

and both U and V are functionally related to X and Y by

$$m_u = g(m_x, m_y)$$
$$m_v = h(m_x, m_y)$$

then U and V are functionally related. We will need the covariance $\sigma_{u\bar{v}} = \rho_{u\bar{v}}\sigma_u\sigma_{\bar{v}}$ to calculate $\sigma^2_{\bar{w}}$. The covariance $\sigma_{u\bar{v}}$ is given approximately by

$$\sigma_{u\bar{v}} = \left[\frac{\partial g}{\partial x} \cdot \frac{\partial h}{\partial x}\right]\sigma_x^2 + \left[\frac{\partial g}{\partial y} \cdot \frac{\partial h}{\partial y}\right]\sigma_y^2 + \left\{\left[\frac{\partial g}{\partial x} \cdot \frac{\partial h}{\partial y}\right] + \left[\frac{\partial g}{\partial y} \cdot \frac{\partial h}{\partial x}\right]\right\}\rho_{xy}\sigma_x\sigma_y \tag{2-13}$$

The square brackets mean, as before, that the partial derivatives are to be evaluated at $\bar{x}$ and $\bar{y}$. If X and Y are independent, the last term again vanishes.

These formulas can be extended to three or more variables if necessary. For convenience, a few special formulas for commonly encountered functions are listed in Table 2-3 with X, Y assumed to be independent. These may be derived from the above formulas as exercises.

In these formulas, if

(a) the partial derivatives when evaluated at the averages are small, and

(b) σ_x, σ_y are small compared to $\bar{x}$, $\bar{y}$,

then the approximations are good and $\bar{w}$ tends to be distributed normally (the ones marked by asterisks are highly skewed and normal approximation could be seriously in error for small n).

Pooling Estimates of Variances. The problem often arises that there are several estimates of a common variance σ^2 which we wish to combine into a single estimate. For example, a gage block may be compared with the master block n_1 times, resulting in an estimate of the variance s_1^2. Another gage block compared with the master block n_2 times, giving rise to s_2^2, etc. As long as the nominal thicknesses of these blocks are within a certain range, the precision of calibration can be expected to remain the same. To get a better evaluation of the precision of the calibration process, we would wish to combine these estimates. The rule is to combine the computed variances weighted by their respective degrees of freedom, or

$$s_p^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2 + \cdots + \nu_k s_k^2}{\nu_1 + \nu_2 + \cdots + \nu_k} \tag{2-14}$$

The pooled estimate of the standard deviation, of course, is $\sqrt{s_p^2} = s_p$. In the example, $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1, \ldots,$ $\nu_k = n_k - 1$, thus the expression reduces to

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{n_1 + n_2 + \cdots + n_k - k} \tag{2-15}$$

The degrees of freedom for the pooled estimate is the sum of the degrees of freedom of individual estimates, or $\nu_1 + \nu_2 + \cdots \nu_k = n_1 + n_2 + \cdots + n_k - k$. With the increased number of degrees of freedom, s_p is a more dependable estimate of σ than an individual s. Eventually, we may consider the value of s_p to be equal to that of σ and claim that we know the precision of the measuring process.

For the special case where k sets of duplicate measurements are available, the above formula reduces to:

$$s_p^2 = \frac{1}{2k} \sum_{1}^{k} d_i^2 \tag{2-16}$$

where d_i = difference of duplicate readings. The pooled standard deviation s_p has k degrees of freedom.

For sets of normally distributed measurements where the number of measurements in each set is small, say less than ten, an estimate of the standard deviation can be obtained by multiplying the range of these measurements by a constant. Table 2-4 lists these constants corresponding to the number n of measurements in the set. For large n, considerable information is lost and this procedure is not recommended.

TABLE 2-4. ESTIMATE OF σ FROM THE RANGE*

n	*Multiplying factor*
2	0.886
3	0.591
4	0.486
5	0.430
6	0.395
7	0.370
8	0.351
9	0.337
10	0.325

*Adapted from *Biometrika Tables for Statisticians*, Vol. I, edited by E. S. Pearson and H. O. Hartley, The University Press, Cambridge, 1958.

If there are k sets of n measurements each, the average range $\bar{R}$ can be computed. The standard deviation can be estimated by multiplying the average range by the factor for n.

Component of Variance Between Groups. In pooling estimates of variances from a number of subgroups, we have increased confidence in the value of the estimate obtained. Let us call this estimate the within-group standard deviation, σ_w. The within-group standard deviation σ_w is a proper measure of dispersions of values within the same group, but not necessarily the proper one for dispersions of values belonging to different groups.

If in making calibrations there is a difference between groups, say from day to day, or from set to set, then the limiting means of the groups are not equal. These limiting means may be thought of as individual measurements; thus, it could be assumed that the average of these limiting means will approach a limit which can be called the limiting mean for all the groups. In estimating σ_w^2, the differences of individuals from the respective group means are used. Obviously σ_w does not include the differences between groups. Let us use σ_b^2 to denote the variance corresponding to the differences between groups, i.e., the measure of dispersion of the limiting means of the respective groups about the limiting mean for all groups.

Thus for each individual measurement x, the variance of X has two components, and

$$\sigma^2 = \sigma_b^2 + \sigma_w^2$$

For the group mean $\bar{x}$ with n measurements in the group,

$$\sigma_{\bar{x}}^2 = \sigma_b^2 + \frac{\sigma_w^2}{n}$$

If k groups of n measurements are available giving averages $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, then an estimate of $\sigma_{\bar{x}}^2$ is

$$s_{\bar{x}}^2 = \frac{1}{k-1} \sum_{i=1}^{k} (\bar{x}_i - \bar{\bar{x}})^2$$

with $k - 1$ degrees of freedom, where $\bar{\bar{x}}$ is the average of all nk measurements.

The resolution of the total variance into components attributable to identifiable causes or factors and the estimation of such components of variances are topics treated under analysis of variance and experimental design. For selected treatments and examples see references 5, 6, and 8.

Comparison of Means and Variances

Comparison of means is perhaps one of the most frequently used techniques in metrology. The mean obtained from one measurement process may be compared with a standard value: two series of measurements on the same quantity may be compared; or sets of measurements on more than two quantities may be compared to determine homogeneity of the group of means.

It is to be borne in mind in all of the comparisons discussed below, that we are interested in comparing the limiting means. The sample means and the computed standard errors are used to calculate confidence limits on the difference between two means. The "t" statistic derived from normal distribution theory is used in this procedure since we are assuming either the measurement process is normal, or the sample averages are approximately normally distributed.

Comparison of a Mean with a Standard Value. In calibration of weights at the National Bureau of Standards, the weights to be calibrated are intercompared with sets of standard weights having "accepted" corrections. Accepted corrections are based on years of experience and considered to be exact to the accuracy required. For instance, the accepted correction for the NB'10 gram weight is −0.4040 mg.

The NB'10 is treated as an unknown and calibrated with each set of weights tested using an intercomparison scheme based on a 100-gm standard weight. Hence the observed correction for NB'10 can be computed for each particular calibration. Table 2-5 lists eleven observed corrections of NB'10 during May 1963.

TABLE 2-5. COMPUTATION OF CONFIDENCE LIMITS FOR OBSERVED CORRECTIONS, NB'10 gm*

Date	i	X_i Observed Corrections to standard 10 gm wt in mg
5–1–63	1	−0.4008
5–1–63	2	−0.4053
5–1–63	3	−0.4022
5–2–63	4	−0.4075
5–2–63	5	−0.3994
5–3–63	6	−0.3986
5–6–63	7	−0.4015
5–6–63	8	−0.3992
5–6–63	9	−0.3973
5–7–63	10	−0.4071
5–7–63	11	−0.4012

$\sum x_i = -4.4201$ $\qquad \sum x_i^2 = 1.77623417$

$\bar{x} = -0.40183$ mg $\qquad \frac{(\sum x_i)^2}{n} = 1.77611673$

difference $= 0.00011744$

$$s^2 = \frac{1}{n-1}(0.00011744) = 0.000011744$$

$s = 0.00343$ = computed standard deviation of an observed correction about the mean.

$\frac{s}{\sqrt{n}} = 0.00103$ = computed standard deviation of the mean of eleven corrections.

= computed standard error of the mean.

For a two-sided 95 percent confidence interval for the mean of the above sample of size 11, $\alpha/2 = 0.025$, $\nu = 10$, and the corresponding value of t is equal to 2.228 in the table of t distribution. Therefore,

$$L_l = \bar{x} - t\frac{s}{\sqrt{n}} = -0.40183 - 2.228 \times 0.00103 = -0.40412$$

and

$$L_u = \bar{x} + t\frac{s}{\sqrt{n}} = -0.40183 + 2.228 \times 0.00103 = -0.39954$$

*Data supplied by Robert Raybold, Metrology Division, National Bureau of Standards.

Calculated 95 percent confidence limits from the eleven observed corrections are -0.4041 and -0.3995. These values include the accepted value of -0.4040, and we conclude that the observed corrections agree with the accepted value.

What if the computed confidence limits for the observed correction do not cover the accepted value? Three explanations may be suggested:

1. The accepted value is correct. However, in choosing $\alpha = 0.05$, we know that 5 percent of the time in the long run we will make an error in our statement. By chance alone, it is possible that this particular set of limits would not cover the accepted value.
2. The average of the observed corrections does not agree with the accepted value because of certain systematic error, temporary or seasonal, particular to one or several members of this set of data for which no adjustment has been made.
3. The accepted value is incorrect, e.g., the mass of the standard has changed.

In our example, we would be extremely reluctant to agree to the third explanation since we have much more confidence in the accepted value than the value based only on eleven calibrations. We are warned that something may have gone wrong, but not unduly alarmed since such an event will happen purely by chance about once every twenty times.

The control chart for mean with known value, to be discussed in a following section, would be the proper tool to use to monitor the constancy of the correction of the standard mass.

Comparison Among Two or More Means. The difference between two quantities X and Y to be measured is the quantity

$$m_{x-y} = m_x - m_y$$

and is estimated by $\bar{x} - \bar{y}$, where $\bar{x}$ and $\bar{y}$ are averages of a number of measurements of X and Y respectively.

Suppose we are interested in knowing whether the difference m_{x-y} could be zero. This problem can be solved by a technique previously introduced, i.e., the confidence limits can be computed for m_{x-y}, and if the upper and lower limits include zero, we could conclude that m_{x-y} may take the value zero; otherwise, we conclude that the evidence is against $m_{x-y} = 0$.

Let us assume that measurements of X and Y are independent with known variances σ_x^2 and σ_y^2 respectively.

By Eq. (2.10)

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n} \text{ for } \bar{x} \text{ of } n \text{ measurements}$$

$$\sigma_{\bar{y}}^2 = \frac{\sigma_y^2}{k} \text{ for } \bar{y} \text{ of } k \text{ measurements}$$

then by (2.8),

$$\sigma^2_{\bar{x}-\bar{y}} = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{k}$$

Therefore, the quantity

$$z = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{k}}} \tag{2-17}$$

is approximately normally distributed with mean zero and a standard deviation of one under the assumption $m_{x-y} = 0$.

If σ_x and σ_y are not known, but the two can be assumed to be approximately equal, e.g., $\bar{x}$ and $\bar{y}$ are measured by the same process, then s_x^2 and s_y^2 can be pooled by Eq. (2-15), or

$$s_p^2 = \frac{(n-1)s_x^2 + (k-1)s_y^2}{n + k - 2}$$

This pooled computed variance estimates

$$\sigma^2 = \sigma_x^2 = \sigma_y^2$$

so that

$$\sigma^2_{\bar{x}-\bar{y}} = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{k} = \frac{n+k}{nk}\sigma^2$$

Thus, the quantity

$$t = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{n+k}{nk}}s_p} \tag{2-18}$$

is distributed as Student's "t", and a confidence interval can be set about m_{x-y} with $\nu = n + k - 2$ and $p = 1 - \alpha$. If this interval does not include zero, we may conclude that the evidence is strongly against the hypothesis $m_x = m_y$.

As an example, we continue with the calibration of weights with NB'10 gm. For 11 subsequent observed corrections during September and October, the confidence interval (computed in the same manner as in the preceding example) has been found to be

$$L_l = -0.40782$$

$$L_u = -0.40126$$

Also,

$$\bar{Y} = -0.40454 \quad \text{and} \quad \frac{s}{\sqrt{k}} = 0.00147$$

It is desired to compare the means of observed corrections for the two sets of data. Here

$$n = k = 11$$

$$\bar{x} = -0.40183, \qquad \bar{y} = -0.40454$$

$$s_x^2 = 0.000011669, \qquad s_y^2 = 0.000023813$$

$$s_p^2 = \tfrac{1}{2}(0.000035482) = 0.000017741$$

$$\frac{n+k}{nk} = \frac{11+11}{121} = \frac{2}{11}$$

$$\sqrt{\frac{n+k}{nk}}s_p = \sqrt{\frac{2}{11} \times 0.000017741} = 0.00180$$

For $\alpha/2 = 0.025$, $1 - \alpha = 0.95$, and $\nu = 20$, $t = 2.086$. Therefore,

$$L_u = (\bar{x} - \bar{y}) + t\sqrt{\frac{n+k}{nk}}\,s_p = 0.00271 + 2.086 \times 0.00180$$
$$= 0.00646$$

$$L_l = (\bar{x} - \bar{y}) - t\sqrt{\frac{n+k}{nk}}\,s_p = -0.00104$$

Since $L_l < 0 < L_u$ shows that the confidence interval includes zero, we conclude that there is no evidence against the hypothesis that the two observed average corrections are the same, or $m_x = m_y$. Note, however, that we would reach a conclusion of no difference wherever the magnitude of $\bar{x} - \bar{y}$ (0.00271 mg) is less than the half-width of the confidence interval ($2.086 \times 0.00180 = 0.00375$ mg) calculated for the particular case. When the true difference m_{x-y} is large, the above situation is not likely to happen; but when the true difference is small, say about 0.003 mg, then it is highly probable that a conclusion of no difference will still be reached. If a detection of difference of this magnitude is of interest, more measurements will be needed.

The following additional topics are treated in reference 4.

1. Sample sizes required under certain specified conditions—Tables A-8 and A-9.
2. σ_x^2 cannot be assumed to be equal to σ_y^2—Section 3-3.1.2.
3. Comparison of several means by Studentized range—Sections 3-4 and 15-4.

Comparison of variances or ranges. As we have seen, the precision of a measurement process can be expressed in terms of the computed standard deviation, the variance, or the range. To compare the precision of two processes a and b, any of the three measures can be used, depending on the preference and convenience of the user.

Let s_a^2 be the estimate of σ_a^2 with ν_a degrees of freedom, and s_b^2 be the estimate of σ_b^2 with ν_b degrees of freedom. The ratio $F = s_a^2/s_b^2$ has a distribution depending on ν_a and ν_b. Tables of upper percentage points of F are given in most statistical textbooks, e.g., reference 4, Table A-5 and Section 4-2.

In the comparison of means, we were interested in finding out if the absolute difference between m_a and m_b could reasonably be zero; similarly, here we may be interested in whether $\sigma_a^2 = \sigma_b^2$, or $\sigma_a^2/\sigma_b^2 = 1$. In practice, however, we are usually concerned with whether the imprecision of one

process exceeds that of another process. We could, therefore, compute the ratio of s_a^2 to s_b^2, and the question arises: If in fact $\sigma_a^2 = \sigma_b^2$, what is the probability of getting a value of the ratio as large as the one observed? For each pair of values of ν_a and ν_b, the tables list the values of F which are exceeded with probability α, the upper percentage point of the distribution of F. If the computed value of F exceeds this tabulated value of $F_{\alpha', \nu_a, \nu_b}$, then we conclude that the evidence is against the hypothesis $\sigma_a^2 = \sigma_b^2$; if it is less, we conclude that σ_a^2 could be equal to σ_b^2.

For example, we could compute the ratio of s_y^2 to s_x^2 in the preceding two examples.

Here the degrees of freedom $\nu_y = \nu_x = 10$, the tabulated value of F which is exceeded 5 percent of the time for these degrees of freedom is 2.98, and

$$\frac{s_y^2}{s_x^2} = \frac{0.000023813}{0.000011669} = 2.041$$

Since 2.04 is less than 2.98, we conclude that there is no reason to believe that the precision of the calibration process in September and October is poorer than that of May.

For small degrees of freedom, the critical value of F is rather large, e.g., for $\nu_a = \nu_b = 3$, and $\alpha' = 0.05$, the value of F is 9.28. It follows that a small difference between σ_a^2 and σ_b^2 is not likely to be detected with a small number of measurements from each process. The table below gives the approximate number of measurements required to have a four-out-of-five chance of detecting whether σ_a is the indicated multiple of σ_b (while maintaining at 0.05 the probability of incorrectly concluding that $\sigma_a > \sigma_b$, when in fact $\sigma_a = \sigma_b$).

Multiple	*No. of measurements*
1.5	39
2.0	15
2.5	9
3.0	7
3.5	6
4.0	5

Table A-11 in reference 4 gives the critical values of the ratios of ranges, and Tables A-20 and A-21 give confidence limits on the standard deviation of the process based on computed standard deviation.

Control Charts Technique for Maintaining Stability and Precision

A laboratory which performs routine measurement or calibration operations yields, as its daily product, numbers—averages, standard deviations, and ranges. The control chart techniques therefore could be applied to these

numbers as products of a manufacturing process to furnish graphical evidence on whether the measurement process is in statistical control or out of statistical control. If it is out of control, these charts usually also indicate where and when the trouble occurred.

Control Chart for Averages. The basic concept of a control chart is in accord with what has been discussed thus far. A measurement process with limiting mean m and standard deviation σ is assumed. The sequence of numbers produced is divided into "rational" subgroups, e.g., by day, by a set of calibrations, etc. The averages of these subgroups are computed. These averages will have a mean m and a standard deviation $\sigma/\sqrt{n}$ where n is the number of measurements within each subgroup. These averages are approximately normally distributed.

In the construction of the control chart for averages, m is plotted as the center line, $m + k(\sigma/\sqrt{n})$ and $m - k(\sigma/\sqrt{n})$ are plotted as control limits, and the averages are plotted in an orderly sequence. If k is taken to be 3, we know that the chance of a plotted point falling outside of the limits, if the process is in control, is very small. Therefore, if a plotted point falls outside these limits, a warning is sounded and investigative action to locate the "assignable" cause that produced the departure, or corrective measures, are called for.

The above reasoning would be applicable to actual cases only if we have chosen the proper standard deviation σ. If the standard deviation is estimated by pooling the estimates computed from each subgroup and denoted by σ_w (within group), obviously differences, if any, between group averages have not been taken into consideration. Where there are between-group differences the variance of the individual $\bar{x}$ is not σ_w^2/n, but, as we have seen before, $\sigma_b^2 + (\sigma_w^2/n)$, where σ_b^2 represents the variance due to differences between groups. If σ_b^2 is of any consequence as compared to σ_w^2, many of the $\bar{x}$ values would exceed the limits constructed by using σ_w alone.

Two alternatives are open to us: (1) remove the cause of the between-group variation; or, (2) if such variation is a proper component of error, take it into account as has been previously discussed.

As an illustration of the use of a control chart on averages, we use again the NB'10 gram data. One hundred observed corrections for NB'10 are plotted in Fig. 2-5, including the two sets of data given under comparison of means (points 18 through 28, and points 60 through 71). A three-sigma limit of 8.6 μg was used based on the "accepted" value of standard deviation.

We note that all the averages are within the control limits, excepting numbers 36, 47, 63, 85, and 87. Five in a hundred falling outside of the three-sigma limits is more than predicted by the theory. No particular reasons, however, could be found for these departures.

Since the accepted value of the standard deviation was obtained by pooling a large number of computed standard deviations for within-sets of

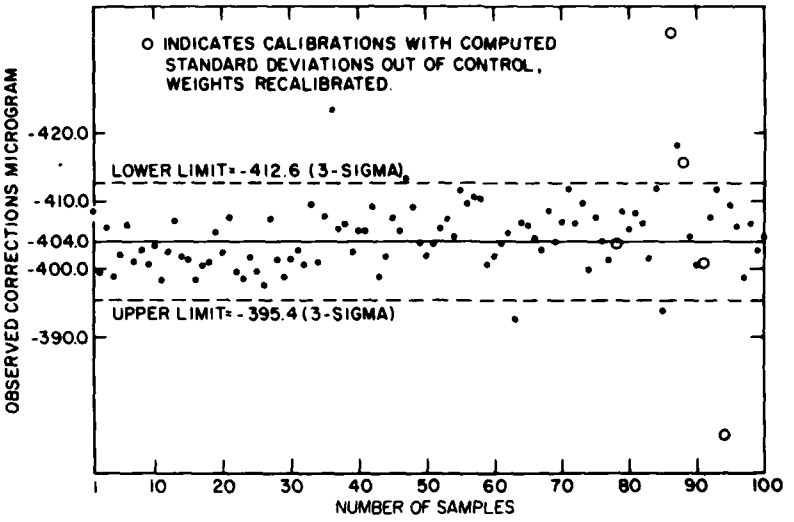


Fig. 2-5. Control chart on $\bar{x}$ for NB'10 gram.

calibrations, the graph indicates that a "between-set" component may be present. A slight shift upwards is also noted between the first 30 points and the remainder.

Control Chart for Standard Deviations. The computed standard deviation, as previously stated, is a measure of imprecision. For a set of calibrations, however, the number of measurements is usually small, and consequently also the degrees of freedom. These computed standard deviations with few degrees of freedom can vary considerably by chance alone, even though the precision of the process remains unchanged. The control chart on the computed standard deviations (or ranges) is therefore an indispensable tool.

The distribution of s depends on the degrees of freedom associated with it, and is not symmetrical about m_s. The frequency curve of s is limited on the left side by zero, and has a long "tail" to the right. The limits, therefore, are not symmetrical about m_s. Furthermore, if the standard deviation of the process is known to be σ, m_s is not equal to σ, but is equal to $c_2\sigma$, where c_2 is a constant associated with the degrees of freedom in s.

The constants necessary for the construction of three-sigma control limits for averages, computed standard deviations, and ranges, are given in most textbooks on quality control. Section 18-3 of reference 4 gives such a table. A more comprehensive treatment on control charts is given in ASTM "Manual on Quality Control of Materials," Special Technical Publication 15-C.

Unfortunately, the notation employed in quality control work differs

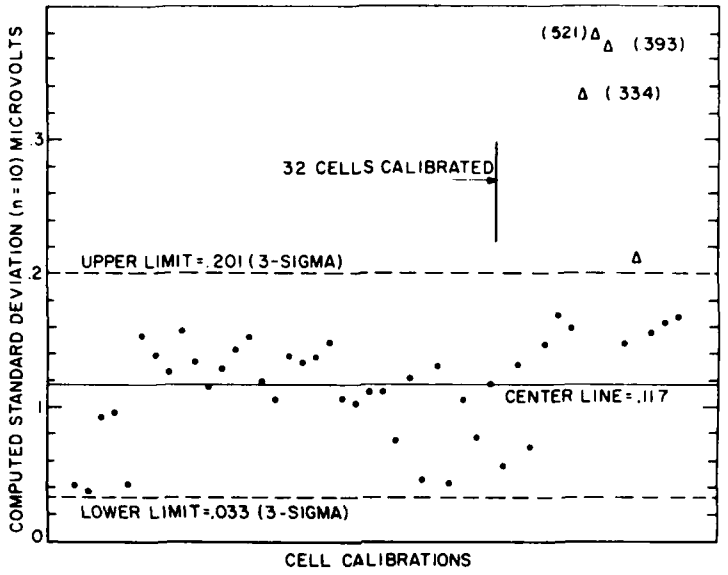


Fig. 2-6. Control chart on s for the calibration of standard cells.

in some respect from what is now standard in statistics, and correction factors have to be applied to some of these constants when the computed standard deviation is calculated by the definition given in this chapter. These corrections are explained in the footnote under the table.

As an example of the use of control charts on the precision of a calibration process, we will use data from NBS calibration of standard cells.* Standard cells in groups of four or six are usually compared with an NBS standard cell on ten separate days. A typical data sheet for a group of six cells, after all the necessary corrections, appears in Table 2-6. The standard deviation of a comparison is calculated from the ten comparisons for each cell and the standard deviation for the average value of the ten comparisons is listed in the line marked SDA. These values were plotted as points 6 through 11 in Fig. 2-6.

Let us assume that the precision of the calibration process remains the same. We can therefore pool the standard deviations computed for each cell (with nine degrees of freedom) over a number of cells and take this value as the current value of the standard deviation of a comparison, σ. The corresponding current value of standard deviation of the average of ten comparisons will be denoted by $\sigma' = \sigma/\sqrt{10}$. The control chart will be made on $s' = s/\sqrt{10}$.

*Illustrative data supplied by Miss Catherine Law, Electricity Division, National Bureau of Standards.

For example, the SDA's for 32 cells calibrated between June 29 and August 8, 1962, are plotted as the first 32 points in Fig. 2-6. The pooled standard deviation of the average is 0.114 with 288 degrees of freedom. The between-group component is assumed to be negligible.

TABLE 2-6. CALIBRATION DATA FOR SIX STANDARD CELLS

Day	Corrected Emf's and standard deviations, Microvolts					
1	27.10	24.30	31.30	33.30	32.30	23.20
2	25.96	24.06	31.06	34.16	33.26	23.76
3	26.02	24.22	31.92	33.82	33.22	24.02
4	26.26	24.96	31.26	33.96	33.26	24.16
5	27.23	25.23	31.53	34.73	33.33	24.43
6	25.90	24.40	31.80	33.90	32.90	24.10
7	26.79	24.99	32.19	34.39	33.39	24.39
8	26.18	24.98	32.18	35.08	33.98	24.38
9	26.17	25.07	31.97	34.27	33.07	23.97
10	26.16	25.16	31.96	34.06	32.96	24.16
R	1.331	1.169	1.127	1.777	1.677	1.233
AVG	26.378	24.738	31.718	34.168	33.168	24.058
SD	0.482	0.439	0.402	0.495	0.425	0.366
SDA	0.153	0.139	0.127	0.157	0.134	0.116

Position	Emf, volts	Position	Emf, volts
1	1.0182264	4	1.0182342
2	1.0182247	5	1.0182332
3	1.0182317	6	1.0182240

Since $n = 10$, we find our constants for three-sigma control limits on s' in Section 18-3 of reference 4 and apply the corrections as follows:

$$\text{Center line} = \sqrt{\frac{n}{n-1}}\, c_2\sigma' = 1.111 \times 0.9227 \times 0.114 = 0.117$$

$$\text{Lower limit} = \sqrt{\frac{n}{n-1}}\, B_1\sigma' = 1.111 \times 0.262 \times 0.114 = 0.033$$

$$\text{Upper limit} = \sqrt{\frac{n}{n-1}}\, B_2\sigma' = 1.111 \times 1.584 \times 0.114 = 0.201$$

The control chart (Fig. 2-6) was constructed using these values of center line and control limits computed from the 32 calibrations. The standard deviations of the averages of subsequent calibrations are then plotted.

Three points in Fig. 2-6 far exceed the upper control limit. All three cells, which were from the same source, showed drifts during the period of calibration. A fourth point barely exceeded the limit. It is to be noted that the data here were selected to include these three points for purposes of illustration only, and do not represent the normal sequence of calibrations.

The main function of the chart is to justify the precision statement on the report of calibration, which is based on a value of σ estimated with

perhaps thousands of degrees of freedom and which is shown to be in control. The report of calibration for these cells ($\sigma = 0.117 \doteq 0.12$) could read:

> "Each value is the mean of ten observations made between ____ and ____. Based on a standard deviation of 0.12 microvolts for the means, these values are correct to 0.36 microvolts relative to the volt as maintained by the national reference group."

Linear Relationship and Fitting of Constants by Least Squares

In using the arithmetic mean of n measurements as an estimate of the limiting mean, we have, knowingly or unknowingly, fitted a constant to the data by the method of least squares, i.e., we have selected a value $\hat{m}$ for m such that

$$\sum_1^n (y_i - \hat{m})^2 = \sum_1^n d_i^2$$

is a minimum. The solution is $\hat{m} = \bar{y}$. The deviations $d_i = y_i - \hat{m} = y_i - \bar{y}$ are called residuals.

Here we can express our measurements in the form of a mathematical model

$$Y = m + \epsilon \tag{2-19}$$

where Y stands for the observed values, m the limiting mean (a constant), and ϵ the random error (normal) of measurement with a limiting mean zero and a standard deviation σ. By (2-1) and (2-9), it follows that

$$m_y = m + m_\epsilon = m$$

and

$$\sigma_y^2 = \sigma^2$$

The method of least squares requires us to use that estimator $\hat{m}$ for m such that the sum of squares of the residuals is a minimum (among all possible estimators). As a corollary, the method also states that the sum of squares of residuals divided by the number of measurements n less the number of estimated constants p will give us an estimate of σ^2, i.e.,

$$s^2 = \frac{\sum (y_i - \hat{m})^2}{n - p} = \frac{\sum (y_i - \bar{y})^2}{n - 1} \tag{2-20}$$

It is seen that the above agrees with our definition of s^2.

Suppose Y, the quantity measured, exhibits a linear functional relationship with a variable which can be controlled accurately; then a model can be written as

$$Y = a + bX + \epsilon \tag{2-21}$$

where, as before, Y is the quantity measured, a (the intercept) and b (the

slope) are two constants to be estimated, and ϵ the random error with limiting mean zero and variance σ^2. We set X at x_i, and observe y_i. For example, y_i might be the change in length of a gage block steel observed for n equally spaced temperatures x_i within a certain range. The quantity of interest is the coefficient of thermal expansion b.

For any estimates of a and b, say $\hat{a}$ and $\hat{b}$, we can compute a value $\hat{y}_i$ for each x_i, or

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

If we require the sum of squares of the residuals

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

to be a minimum, then it can be shown that

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2-22}$$

and

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{2-23}$$

The variance of Y can be estimated by

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2} \tag{2-24}$$

with $n - 2$ degrees of freedom since two constants have been estimated from the data.

The standard errors of $\hat{b}$ and $\hat{a}$ are respectively estimated by $s_{\hat{b}}$ and $s_{\hat{a}}$, where

$$s_{\hat{b}}^2 = \frac{s^2}{\sum (x_i - \bar{x})^2} \tag{2-25}$$

$$s_{\hat{a}}^2 = s^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right] \tag{2-26}$$

With these estimates and the degrees of freedom associated with s^2, confidence limits can be computed for $\hat{a}$ and $\hat{b}$ for the confidence coefficient selected if we assume that errors are normally distributed.

Thus, the lower and upper limits of a and b, respectively, are:

$$\hat{a} - ts_{\hat{a}}, \qquad \hat{a} + ts_{\hat{a}}$$

$$\hat{b} - ts_{\hat{b}}, \qquad \hat{b} + ts_{\hat{b}}$$

for the value of t corresponding to the degree of freedom and the selected confidence coefficient.

The following problems relating to a linear relationship between two variables are treated in reference 4, Section 5-4.

1. Confidence intervals for a point on the fitted line.
2. Confidence band for the line as a whole.
3. Confidence interval for a single predicted value of Y for a given X.

Polynomial and multivariate relationships are treated in Chapter 6 of the same reference.

REFERENCES

The following references are recommended for topics introduced in the first section of this chapter:

1. Wilson, Jr., E. B., *An Introduction to Scientific Research*, McGraw-Hill Book Company, New York, 1952, Chapters 7, 8, and 9.
2. Eisenhart, Churchill, "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration System," *Journal of Research of the National Bureau of Standards*, Vol. 67C, No. 2, 1963.
3. Youden, W. J., *Experimentation and Measurement*, National Science Teacher Association Vista of Science Series No. 2, Scholastic Book Series, New York.

In addition to the three general references given above, the following are selected with special emphasis on their ease of understanding and applicability in the measurement science:

Statistical Methods

4. Natrella, M. G., *Experimental Statistics*, NBS Handbook 91, U.S. Government Printing Office, Washington, D.C., 1963.
5. Youden, W. J., *Statistical Methods for Chemists*, John Wiley & Sons, Inc., New York, 1951.
6. Davies, O. L., *Statistical Method in Research and Production* (3rd ed.), Hafner Publishing Co., Inc., New York, 1957.

Textbooks

7. Dixon, W. J. and F. J. Massey, *Introduction to Statistical Analysis* (2nd ed.), McGraw-Hill Book Company, New York, 1957.
8. Brownlee, K. A., *Statistical Theory and Methodology in Science and Engineering*, John Wiley & Sons, Inc., New York, 1960.
9. Areley, N. and K. R. Buch, *Introduction to the Theory of Probability and Statistics*, John Wiley & Sons, Inc., New York, 1950.

JOURNAL OF RESEARCH of the National Bureau of Standards – C. Engineering and Instrumentation
Vol. 70C, No. 4, October–December 1966

Notes on the Use of Propagation of Error Formulas

H. H. Ku

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(May 27, 1966)

The "law of propagation of error" is a tool that physical scientists have conveniently and frequently used in their work for many years, yet an adequate reference is difficult to find. In this paper an expository review of this topic is presented, particularly in the light of current practices and interpretations. Examples on the accuracy of the approximations are given. The reporting of the uncertainties of final results is discussed.

Key Words: Approximation, error, formula, imprecision, law of error, products, propagation of error, random, ratio, systematic, sum.

Introduction

In the December 1939, issue of the American Physics Teacher, Raymond T. Birge wrote an expository paper on "The Propagation of Errors." In the introductory paragraph of his paper, Birge remarked:

"The question of what constitutes the most reliable value to be assigned as the uncertainty of any given measured quantity is one that has been discussed for many decades and, presumably, will continue to be discussed. It is a question that involves many considerations and by its very nature has no unique answer. The subject of the propagation of errors, on the contrary, is a purely mathematical matter, with very definite and easily ascertained conclusions. Although th general subject of the present article is by no means new,[1] many scientists still fail to avail themselves of the enlightening conclusions that may often thus be reached, while others frequently use the theory incorrectly and thus arrive at quite misleading conclusions."

Birge's remark 27 years ago still sounds fitting today. For a number of years, the need for an expository paper on this topic has been felt by the staff of the Statistical Engineering Laboratory at the National Bureau of Standards. Frequent inquiries have to be answered, yet a diligent search in current literature and textbooks failed to produce a suitable reference that treats the subject matter adequately. The present manuscript was written to fill this need.

In section 1, we consider the two distinct situations under which the propagation of error formulas can be used. The mathematical manipulations are the same, yet the interpretations of the results are entirely different. In section 2 the notations are defined and the general formulas given. Frequently used special formulas are listed at the end of the section for convenient reference. In section 3 the accuracies of the approximations are discussed, together with suggestions on the use of the errors propagated. Section 4 contains suggestions on the reporting of final results.

[1] See, for instance, M. Merriman, Method of Least Squares, pp. 75–79 (ed. 8, 1910).

The "law of propagation of error" is a tool that physical scientists have conveniently and frequently used in their work for many years. No claim is made here that it is the only tool or even a suitable tool for all occasions. "Data analysis" is an ever-expanding field and other methods, existing or new, are probably available for the analysis and interpretation for each particular set of data. Nevertheless, under certain assumptions given in detail in the following sections, the approximations resulting from the use of these formulas are useful in giving an estimate of the uncertainty of a reported value. The uncertainty computed from the use of these formulas, however, is probably somewhat less that the actual in the sense that no function form is known exactly and the number of variables considered usually does not represent fully the contributors of errors that affect the final result.

1. Statistical Tolerancing Versus Imprecision of a Derived Quantity

1.1. Propagation of error formulas are frequently used by engineers in the type of problem called "Statistical tolerancing." In such problems, we are concerned with the behavior of the characteristic W of a system as related to the behavior of a characteristic X of its component. For instance, an engineer may have designed a circuit. A property W of the circuit may be related to the value X of the resistance used. As the value of X is changed, W changes and the relationship can be expressed by a mathematical function

$$W = F(X)$$

within a certain range of the values of X.

Suppose our engineer decides on $W = w_0$ to be the desired property of the circuit, and specifies $X = x_0$ for this purpose. He realizes, however, that there

will be variations among the large lot of resistors he ordered, no matter how tight his specifications are. Let x denote the value of any one of the resistors in the lot, then some of the time x will be below x_0, while at other times x will be above x_0. In other words, x has a distribution of values somewhat clustered about x_0. As x varies with each resistor, so does w with each circuit manufactured.

If our engineer knows the mean and standard deviation (or variance) of x, based on data from the history of their manufacture, then he can calculate the approximate mean and variance of w by the propagation of error formulas:

$$\text{mean } (w) \doteq F(\text{mean } x), \text{ and}$$

$$\text{variance } (w) \doteq \left[\frac{dF}{dX}\right]^2 \text{var } (x), \tag{1.1}$$

where the square brackets signify that the derivatives within the brackets are to be evaluated at the mean of x. The approximations computed refer to the mean and variance of an individual unit in the collection of circuits that will be manufactured from the lot of resistors. The distribution of values of w, however, is still far from being determined since it depends entirely on the functional form of the relation between W and X, as mathematical variables, and the distribution of x itself, as a random variable. This type of approach has been used frequently in preliminary examinations of the reliability of performance of a system, where X may be considered as a multidimensional variable.

1.2 Let us consider now the second situation under which propagation of error formulas are used. This situation is the one considered in Birge's paper, and is the one that will be discussed in the main part of this paper.

A physicist may wish to determine the "true" value w_0 of interest, for example, the atomic weight of silver. He makes n independent measurements on some related quantity x and calculates

$$\bar{x}_n = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

as an estimate of the true value x_0 and

$$s^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x}_n)^2$$

as an index of dispersion of his measured values. The physicist is mainly concerned in obtaining an estimate $\hat{w}$ of w_0, and of the standard deviation of $\hat{w}$ as a measure of precision of his result. He therefore computes by the propagation of error formulas:

$$\hat{w} = F(\bar{x}_n)$$

$$\hat{\text{var}}\ (\hat{w}) = \left[\frac{dF}{dX}\right]^2 \frac{s^2}{n} \tag{1.2}$$

$$\hat{\sigma}_{\hat{w}} = \sqrt{\hat{\text{var}}\ (\hat{w})}$$

Often he assumes that $\hat{w}$ is distributed at least approximately in accordance with the normal law of error and gives probability limits to the statistical uncertainty of his estimate $\hat{w}$ based on the standard deviation calculated ($\hat{\sigma}_{\hat{w}}$) and this assumption.

Cramér [1946] has shown that under very general conditions, functions of sample moments are asymptotically normal, with mean and variance given by the respective propagation of error formulas.[2] Since $\bar{x}_n$ is the first sample moment, the estimate $\hat{w}$ will be approximately normally distributed for large n. Hence our physicist is interested in the variance (or the standard deviation) of the normal distribution which the distribution of $F(\bar{x}_n)$ approximates as n increases. (Note that both estimators $\hat{w}$ and $\hat{\text{var}}\ (\hat{w})$ are functions of n.) For n large, the distribution of $\hat{w}$ can be assumed to be approximately normal and probability statements can be made about $\hat{w}$.

1.3 Hence, we have the two cases:

(1) The problem of determining the mean and variance (or standard deviation) of the actual distribution of a given function $F(x)$ of a particular random variable x, and

(2) The problem of estimating the mean and variance (or standard deviation) of the normal distribution to which the distribution of $F(\bar{x}_n)$ tends asymptotically.

As examples of problems studied under the first case, we can cite Fieller [1932] on the ratio of two normally distributed random variables, and Craig [1937] and Goodman [1962] on the product of two or more random variables. Tukey, in three Princeton University reports, extended the classical formulas through the fourth order terms for the mean and variance, and propagated the skewness and elongation of the distribution of $F(x)$ as well. These reports present perhaps the most exhaustive treatment of statistical tolerancing to date.

From now on we shall be concerned in this paper with the second case only, i.e., the problem of estimating the mean and variance, or standard deviation, of the normal distribution to which the distribution of $F(\bar{x}_n)$ tends as n increases indefinitely, and hence also the problem of using approximations to the mean and variance computed from a finite number of measurements. Since the mean and standard deviation are the parameters that specify a particular normal distribution, our problem is by its very nature less complicated than that of statistical tolerancing where the actual distribution of the function may have to be specified. We shall, however, utilize formulas given in Tukey's reports to check on the adequacy of some of the approximations.

[2] A brief summary is given in paragraph 2.2

2. Propagation of Error Formulas

2.1. Definitions and Notations

(1) X, Y, Z in capitals stand for the mathematical variables to be measured; x, y, z in lower cases stand for the measured values of these variables: x_i, y_j, z_k with subscripts stand for the particular values of the ith measurement on x, the jth on y, and the kth on z, respectively.

(2) $W = f(X, Y, Z)$ is a continuous function of the variables X, Y, Z, with derivatives

$$\frac{\partial W}{\partial X}, \frac{\partial^2 W}{\partial X \partial Y}, \text{ etc.}$$

(3) All derivatives appearing in square brackets, for example $\left[\frac{\partial W}{\partial X}\right], \left[\frac{\partial W}{\partial Y}\right]$, stand for the values of these derivatives evaluated at the means of x and y, if known, or at the sample averages of x and y, if the means are not known.

(4) In order to emphasize the fact that the mean M, variance σ^2 and other population parameters are usually not known, we list here symbols for both the estimators of population values and the population values. For a particular set of values of x, the values computed from these estimators are estimates, or computed values of these estimators.

Estimators of parameters	Corresponding population parameters
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	M_x (mean = first moment)
$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	σ_x^2 (variance = second central moment)
$= \frac{1}{n-1}\left\{\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right\}$	
$s_{xy} = s_{yx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	$\sigma_{xy} = \sigma_{yx}$ (covariance)
$= \frac{1}{n-1}\left\{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right\}$	
$r_{xy} = \frac{s_{xy}}{s_x s_y} = r_{yx}$	ρ_{xy} (correlation coefficient)
s_x	σ_x (standard deviation of x about M_x)
$s_{\bar{x}} = \frac{1}{\sqrt{n}} s_x$	$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ (standard deviation of the average $\bar{x}$, or standard error)
$v_x = \frac{s_x}{\bar{x}}$	$\frac{\sigma_x}{M_x}$ (coefficient of variation or relative standard deviation)

In addition, we use $|\Delta x|$ to denote the bound for possible systematic errors on the measurements of x. The bound of these errors, unknown in sign, is usually established or conjectured by the experimenter and its value is not based on the measurements in hand.

2.2. General Theorem and Remarks

As mentioned briefly in paragraph 1.2, the propagation of error formulas are special applications of results obtained in the study of properties of distributions of functions of sample moments. Doob [1935], Hsu [1949], and others have investigated the limiting distribution of functions of sample means relating to hypothesis testing. Curtiss [1943] derived the limiting means and variances of the several functions of variables in connection with transformations used in the analysis of variance. Cramér, in chapters 27 and 28 of his classical treatise, proved two theorems and also discussed the asymptotic properties of distributions of functions of sample moments in detail. For convenient reference we shall phrase his theorems and remarks in terms of functions of sample averages, to serve as a basis of justification for the use of propagation of error formulas.

THEOREM (Cramér, pp. 366, 352–356)

If, in some neighborhood of the point $X = M_x$, $Y = M_y$, *the function* $F(X, Y)$ *is continuous and has continuous derivatives of the first and second order with respect to the arguments* X *and* Y, *the random variable* $\hat{w} = F(\bar{x}, \bar{y})$ *is asymptotically normal, the mean and variance of the limiting normal distribution being given by:*

$$\text{mean } \hat{w} = F(M_x, M_y) \tag{2.1}$$

$$\text{var } \hat{w} = \left[\frac{\partial F}{\partial X}\right]^2 \frac{\sigma_x^2}{n} + \left[\frac{\partial F}{\partial Y}\right]^2 \frac{\sigma_y^2}{n} + 2\left[\frac{\partial F}{\partial X}\right]\left[\frac{\partial F}{\partial Y}\right]\frac{\sigma_{xy}}{n}. \tag{2.2}$$

REMARK 1. (Cramér, p. 367)

It follows from this theorem that any function of sample averages is, for large values of n, approximately normally distributed about the value of the function determined by the mean values of the basic variables, with a variance of the form C/n, provided only that expressions (2.1) and (2.2) yield finite values for the mean and the variance of the limiting distribution.

REMARK 2. (Cramér, pp. 367, 415, also Doob, Hsu)

In general, the constant C in the expression of the variance will have a positive value. However, in exceptional cases C may be zero, which implies that the variance is of a smaller order than n^{-1}. Then some expression of the form

$$n^p\{\hat{w} - F(M_x, M_y)\}, \qquad p > \tfrac{1}{2},$$

may have a definite limiting distribution, but this is not necessarily normal.

REMARK 3. (Cramér, pp. 366, 213–214)

The function $F(\bar{x}, \bar{y})$ may be asymptotically normal even though the mean and variance of $F(\bar{x}, \bar{y})$ do not

exist, or do not tend to the mean and variance of the limiting normal form. Generally, if the distribution of a random variable w depends on a parameter n, and if two quantities M and σ can be found such that the distribution function of the variable $\frac{w-M}{\sigma}$ tends to $\Phi(t)$ (normal distribution function with mean zero and standard deviation one) as $n \to \infty$, we shall say that w is asymptotically normal (M, σ). This does not imply that the mean and the standard deviation of w tend to M and σ, nor even that these moments exist, but is simply equivalent to saying that for any interval (a, b) not depending on n,

$$\lim_{n \to \infty} \text{Prob.}\ (M + a\sigma < w < M + b\sigma) = \Phi(b) - \Phi(a).$$

EXAMPLE: If x is from a continuous distribution with positive mean and a *finite variance but with positive* probability that some x can take negative values, then the function $\ln \bar{x}$ is not even defined for all values of $\bar{x}$, and therefore the mean of the function $\ln \bar{x}$ does not exist: yet where the mean of $\bar{x}$ has a positive value, (2.1) and (2.2) give the mean and variance of the limiting normal distribution.

2.3. Propagation of Error Formulas

Fortified with the general theorem stated in the preceding paragraph, we shall proceed to derive the traditional propagation of error formulas *in an elementary manner*, making some comments and assumptions that may be of interest. It will be helpful, however, to explain first what is meant here by the term "random error" in a measurement process.

a. Random Errors

In a measurement situation, we consider random errors typically to be the sum total of all the small negligible independent errors over which we have no control—interpolation in reading scales, slight fluctuation *in environmental conditions*, imperfection and nonconstancy of our senses, etc. Thus for a *stable* measurement process, we find that:

(1) The measured values *do follow a distribution*, with small errors occurring more frequently than larger ones, and with positive and negative errors about *balancing one another, and*

(2) there is no obvious trend or pattern in the sequence of measurements.

Let us denote the *i*th measurement of x to be

$$x_i = M_x + \epsilon_i$$

where M_x is the mean of all measurements for the measurement process, and ϵ_i the random error of measurement x_i. Then for condition 1, we assume A_1: The distribution of errors is symmetrical and bell-shaped, with mean zero and standard deviation σ_x, or

$$\begin{aligned} \text{mean } \epsilon_i &= 0 \\ \text{mean } x_i &= M_x \\ \text{var } x &= \text{var } \epsilon_i \\ &= \text{mean } \epsilon_i^2 = \sigma_x^2. \end{aligned}$$

And for condition 2, we assume

A_2: The errors in the measurements of x_i $(i = 1, 2, \ldots, n)$ are statistically independent: in particular these errors are not correlated or associated in any way, i.e.,

$$\text{mean } (\epsilon_i \cdot \epsilon_j) = 0, \qquad i \neq j.$$

Thus for $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$, the mean of $\bar{x}$ is M_x. Furthermore,

$$\bar{x} - M_x = \frac{\epsilon_1 + \epsilon_2 + \ldots + \epsilon_n}{n}.$$

By definition, the variance of $\bar{x}$ is

$$\text{mean } (\bar{x} - M_x)^2 = \text{mean } \left(\frac{\epsilon_1 + \epsilon_2 + \ldots + \epsilon_n}{n}\right)^2$$

$$= \frac{1}{n^2}\left\{\text{mean } \left(\sum_{i=1}^{n} \epsilon_i^2\right) + \text{mean } \left(\sum_{i \neq j} \epsilon_i \epsilon_j\right)\right\}$$

$$= \frac{1}{n^2}\left\{n \text{ mean } (\epsilon_i)^2 + \sum_{i \neq j} \text{mean } (\epsilon_i \epsilon_j)\right\}.$$

Using assumptions A_1 and A_2, we obtain

$$\text{var } (\bar{x}) = \frac{1}{n}\sigma_x^2$$

or the variance of the average of n *independent* measurements is $\frac{1}{n}$ of the variance of an individual measurement.[3]

Here the average $\bar{x}$ is a linear function of the individual x's, and the exact expressions of mean and variance of an average in terms of that of the individual values are well known. For functions that are not linear in the x's, we expand the function about the mean of x by the Taylor series, and assume that the function in the neighborhood of the mean can be approximated by the lower order terms. For example, let

$$\begin{aligned} W &= F(X, Y), \\ x &= M_x + \epsilon_x, \\ y &= M_y + \epsilon_y, \end{aligned}$$

[3] If, however, the measurements are not independent, then this formula is incorrect since the means of products $(\epsilon_i \epsilon_j)$ are not equal to zero. In that case let

$$\text{mean } (\epsilon_i \epsilon_j) = \rho_{ij}\sigma_x^2, \text{ and } \rho = \sum_{i \neq j} \rho_{ij}/n(n-1), \text{ then var } (\bar{x}) = \frac{\sigma_x^2}{n}\{1 + (n-1)\rho\}$$

where each of ϵ_x and ϵ_y satisfies assumptions A_1 and A_2, then we can write

$$F(x, y) = F(M_x, M_y) + \left[\frac{\partial F}{\partial X}\right]\epsilon_x + \left[\frac{\partial F}{\partial Y}\right]\epsilon_y + \frac{1}{2!}\left\{\left[\frac{\partial^2 F}{\partial X^2}\right]\epsilon_x^2 + 2\left[\frac{\partial^2 F}{\partial X \partial Y}\right]\epsilon_x\epsilon_y + \left[\frac{\partial^2 F}{\partial Y^2}\right]\epsilon_y^2\right\} + \text{terms of higher orders in } \epsilon_x \text{ and } \epsilon_y. \quad (2.3)$$

Or, neglecting terms of higher order than ϵ_x^2 and ϵ_y^2,

$$F(x, y) - F(M_x, M_y) \doteq \left[\frac{\partial F}{\partial X}\right]\epsilon_x + \left[\frac{\partial F}{\partial Y}\right]\epsilon_y + \frac{1}{2!}\left\{\left[\frac{\partial^2 F}{\partial X^2}\right]\epsilon_x^2 + 2\left[\frac{\partial^2 F}{\partial X \partial Y}\right]\epsilon_x\epsilon_y + \left[\frac{\partial^2 F}{\partial Y^2}\right]\epsilon_y^2\right\}.$$

Since the means of ϵ_x and ϵ_y are 0, if we take the mean on both sides,

$$\text{mean}\{F(x, y) - F(M_x, M_y)\} \doteq \frac{1}{2}\left\{\left[\frac{\partial^2 F}{\partial X^2}\right]\sigma_x^2 + 2\left[\frac{\partial^2 F}{\partial X \partial Y}\right]\sigma_{xy} + \left[\frac{\partial^2 F}{\partial Y^2}\right]\sigma_y^2\right\}. \quad (2.4)$$

Thus the mean of a function of values always differs from the value of a function of means by a quantity represented by (2.4), approximately. If the function of means $F(M_x, M_y)$ is the value of interest, then to approximate $F(M_x, M_y)$ by the mean of $F(x, y)$ would introduce an error, or bias, the magnitude of which depends on the functional form, the variances of and the covariance between x and y. If, however, we use the function of averages, $F(\bar{x}, \bar{y})$, then

$$\textit{mean}\ \hat{w} = \text{mean}\ F(\bar{x}, \bar{y}) \doteq F(M_x, M_y) + \frac{1}{2}\left\{\left[\frac{\partial^2 F}{\partial X^2}\right]\frac{\sigma_x^2}{n} + 2\left[\frac{\partial^2 F}{\partial X \partial Y}\right]\frac{\sigma_{xy}}{n} + \left[\frac{\partial^2 F}{\partial Y^2}\right]\frac{\sigma_y^2}{n}\right\}, \quad (2.5)$$

and the bias is only $1/n$ times that of the mean of the function of individual values. When n becomes large, this bias tends to zero, and (2.1) results.

This bias can be calculated by (2.5) and compared to the standard deviation of $\hat{w}$. In practice, if σ_x and σ_y are small, the bias is often of a magnitude that is negligible.

To propagate the variance, we note that if ϵ_x and ϵ_y are small in the sense that the second and higher order terms in (2.3) can be collectively neglected in comparison to terms involving ϵ_x and ϵ_y only, then

$$F(x, y) - F(M_x, M_y) \doteq \left[\frac{\partial F}{\partial X}\right]\epsilon_x + \left[\frac{\partial F}{\partial Y}\right]\epsilon_y,$$

and the variance of $F(x, y)$ is, approximately,

$$\text{mean}\{F(x, y) - F(M_x, M_y)\}^2 \doteq \text{mean}\left\{\left[\frac{\partial F}{\partial X}\right]\epsilon_x + \left[\frac{\partial F}{\partial Y}\right]\epsilon_y\right\}^2 = \left[\frac{\partial F}{\partial X}\right]^2\sigma_x^2 + \left[\frac{\partial F}{\partial Y}\right]^2\sigma_y^2 + 2\left[\frac{\partial F}{\partial X}\right]\left[\frac{\partial F}{\partial Y}\right]\sigma_{xy}. \quad (2.6)$$

And for $\hat{w} = F(\bar{x}, \bar{y})$, the variance of $\hat{w}$ is

$$\text{var}(\hat{w}) \doteq \left[\frac{\partial F}{\partial X}\right]^2\frac{\sigma_x^2}{n} + \left[\frac{\partial F}{\partial Y}\right]^2\frac{\sigma_y^2}{n} + 2\left[\frac{\partial F}{\partial X}\right]\left[\frac{\partial F}{\partial Y}\right]\frac{\sigma_{xy}}{n} \quad (2.7)$$

the limiting form of which is (2.2).

Finally, if σ_x^2, σ_y^2, and σ_{xy} are not known, we substitute their estimators in formulas (2.5) and (2.7), resulting in:

$$\widehat{\text{mean}}(\hat{w}) \doteq F(M_x, M_y) + \frac{1}{2}\left\{\left[\frac{\partial^2 F}{\partial X^2}\right]\frac{s_x^2}{n} + \left[\frac{\partial^2 F}{\partial Y^2}\right]\frac{s_y^2}{n} + 2\left[\frac{\partial^2 F}{\partial X \partial Y}\right]\frac{s_{xy}}{n}\right\}, \quad (2.8)$$

and

$$\widehat{\text{var}}(\hat{w}) \doteq \left[\frac{\partial F}{\partial X}\right]^2\frac{s_x^2}{n} + \left[\frac{\partial F}{\partial Y}\right]^2\frac{s_y^2}{n} + 2\left[\frac{\partial F}{\partial X}\right]\left[\frac{\partial F}{\partial Y}\right]\frac{s_{xy}}{n}. \quad (2.9)$$

If we assume further that the random errors in measurements of x and y are independent, then $\sigma_{xy} = 0$, and the terms involving σ_{xy} in (2.5), (2.6), and (2.7) vanishes. If this is the case, the terms involving s_{xy} in (2.8) and (2.9) should also be dropped. This reduced version of the formula for independent x and y,

$$\text{var}(\hat{w}) \doteq \left[\frac{\partial F}{\partial X}\right]^2\frac{\sigma_x^2}{n} + \left[\frac{\partial F}{\partial Y}\right]^2\frac{\sigma_y^2}{n}, \quad (2.10)$$

is of the form given in Birge's paper and in other textbooks on statistical analysis of data [Mandel, 1964, pp. 72–76].

For $W = F(X, Y, Z)$, there will be three variance and three covariance terms in (2.5) and (2.7). Extension to more than three variables presents no new problems.

b. Extension to More Than One Function of the Variables

Let

$$U = g(X, Y, Z),$$

and

$$V = h(X, Y, Z).$$

Then in addition to the above formulas, we have

$$\sigma_w \doteq \left[\frac{\partial U}{\partial X}\cdot\frac{\partial U}{\partial X}\right]\sigma_x^2 + \left[\frac{\partial U}{\partial Y}\cdot\frac{\partial U}{\partial Y}\right]\sigma_y^2 + \left[\frac{\partial U}{\partial Z}\cdot\frac{\partial U}{\partial Z}\right]\sigma_z^2$$
$$+ \left\{\left[\frac{\partial U}{\partial X}\cdot\frac{\partial U}{\partial Y}\right] + \left[\frac{\partial U}{\partial Y}\cdot\frac{\partial V}{\partial X}\right]\right\}\rho_{xy}\sigma_x\sigma_y$$
$$+ \left\{\left[\frac{\partial U}{\partial Y}\cdot\frac{\partial U}{\partial Z}\right] + \left[\frac{\partial U}{\partial Z}\cdot\frac{\partial V}{\partial Y}\right]\right\}\rho_{yz}\sigma_y\sigma_z$$
$$+ \left\{\left[\frac{\partial U}{\partial Z}\cdot\frac{\partial U}{\partial X}\right] + \left[\frac{\partial U}{\partial X}\cdot\frac{\partial U}{\partial Z}\right]\right\}\rho_{zx}\sigma_z\sigma_x. \quad (2.11)$$

Expression (2.11) may be convenient to use to get $\sigma(\hat{w})$ where $W = F(U, V)$, and U and V are known functions of X, Y, and Z.

c. Some Frequently Used Formulas

For convenience, a few special formulas for commonly encountered functions are listed in table 1 with x, y assumed to be independent. These may be derived from the above formulas.

2.4. Systematic Errors

By a systematic error we mean a fixed deviation that is inherent in each and every measurement of x in a particular sequence of measurements. If the magnitude and direction of the systematic error are known, a correction can be made such that $M_x = x_0$, or the mean of the sequence of measurements is equal to the value sought after. If the sign of the systematic error is not known and the magnitude of the error can be only estimated to be within some reasonable bound $|\Delta x|$, perhaps by experience or judgment, then M_x is within the limits $x_0 - \Delta x$ and $x_0 + \Delta x$.

For a function of two variables $W = F(X, Y)$ then, a bound $|\Delta w|$ for the systematic error in W is given by:

$$|\Delta w| \doteq \left|\left[\frac{\partial F}{\partial X}\right]\Delta x\right| + \left|\left[\frac{\partial F}{\partial Y}\right]\Delta y\right|. \quad (2.12)$$

assuming, as before, that Δx and Δy are small such that second and higher order terms in Δx and Δy are collectively negligible in the Taylor series expansion. Since ordinarily we do not know the signs of Δx and Δy, we have no choice but to add the absolute values of the two terms together, even though the signs of the values of the partial derivatives evaluated are known. (If the signs of either Δx or Δy is known, this information, of course, should not be ignored.) If these derivatives are evaluated at the point $\bar{x}$ and $\bar{y}$, then the random components of error of $\bar{x}$ and $\bar{y}$ are required to be small so that these derivatives take approximately the same values as when evaluated at x_0 and y_0.

When there are a number of systematic errors to be propagated, one approach is to take $|\Delta w|$ as the square root of the sum of squares of terms on the right-hand side of (2.12), instead of adding together the absolute values of all the terms. This procedure presupposes that some of the systematic errors may be positive and the others negative, and the two classes cancel each other to a certain extent.

The treatment of inaccuracy due to systematic errors of assignable origin but of unknown magnitudes is discussed in detail in section 4.2 of Eisenhart [1963]. Since there is no generally accepted standard method for combining several systematic errors, Eisenhart advised and we quote

"Therefore, anyone who uses one of these methods for the 'combination of errors' should indicate explicitly which of these (or an alternative method) he has used."

Information on the source and magnitude of each contributing elemental systematic error is, of course, also essential.

3. Practical Accuracies at the Various Stages of Approximations

3.1. From the preceding sections we observe that there are three stages of approximations:

(1) In the Taylor series expansion (2.3), terms higher than the first partial derivatives are considered to be negligible.

(2) $\hat{w}$ is approximately normally distributed for large n. Is the normal distribution still a good approximation for small n?

(3) If σ_x^2 and σ_y^2 are known, we obtain $\sigma_{\hat{w}}^2$ from (2.7), and we can use this value to construct a confidence interval[3a] about $\hat{w}$ with the desired level of confidence (approximate) based on normal theory. If σ_x^2 and σ_y^2 are not known, and s_x^2 and s_y^2 are calculated from a small number of measurements, what can we say about $\hat{w}$ using $\hat{\text{var}}(\hat{w})$ calculated from (2.9)?

To get some numerical feeling for the closeness of these approximations, we shall simplify matters by making the following assumptions which do not seem to be too restrictive in measurement situations:

B_1: x and y are normally and independently distributed, with the ratio M/σ not less than 10.[4]

B_2: The functional forms used are the well-behaved ones that do not possess derivatives assuming unreasonably large values when evaluated at the averages of the individual variables.

Thus for linear functions, such as

$$W = AX + BY,$$

the second and higher derivatives vanish, and (2.6) is exact.

The adequacy of these approximations is studied in paragraphs 3.2 and 3.3 below. In paragraph 3.4 sug-

[3a] See Natrella [1963], sec. 1 to 7, also chs. 2 and 3.

[4] For notational convenience, the symbols w, x, y, σ_x, σ_y, etc., are used in this and the subsequent sections. The corresponding symbols for the average could be used by straight substitution.

TABLE 1. *Propagation of error formulas for some simple functions*

Function form of $\hat{w}$ *	Approx. formula for $\hat{\text{var}}(\hat{w})$ (x and y are assumed to be statistically independent)	Term to be added if x and y are correlated, and a reliable estimate of $\sigma_{\bar{x}\bar{y}}$, $s_{\bar{x}\bar{y}}$, can be assumed
$A\bar{x} + B\bar{y}$	$A^2 s_{\bar{x}}^2 + B^2 s_{\bar{y}}^2$	$2AB s_{\bar{x}\bar{y}}$
$\left(\frac{\bar{x}}{\sigma_{\bar{x}}^2} + \frac{\bar{y}}{\sigma_{\bar{y}}^2}\right) \Big/ \left(\frac{1}{\sigma_{\bar{x}}^2} + \frac{1}{\sigma_{\bar{y}}^2}\right)$ **	$1 \Big/ \left(\frac{1}{\sigma_{\bar{x}}^2} + \frac{1}{\sigma_{\bar{y}}^2}\right)$	
$\frac{\bar{x}}{\bar{y}}$	$\left(\frac{\bar{x}}{\bar{y}}\right)^2 \left(\frac{s_{\bar{x}}^2}{\bar{x}^2} + \frac{s_{\bar{y}}^2}{\bar{y}^2}\right)$	$\left(\frac{\bar{x}}{\bar{y}}\right)^2 \left(-2\frac{s_{\bar{x}\bar{y}}}{\bar{x}\bar{y}}\right)$
$\frac{1}{\bar{y}}$	$\frac{s_{\bar{y}}^2}{\bar{y}^4}$	
$\frac{\bar{x}}{\bar{x} + \bar{y}}$	$\left(\frac{\hat{w}}{\bar{x}}\right)^4 (\bar{y}^2 s_{\bar{x}}^2 + \bar{x}^2 s_{\bar{y}}^2)$	$\left(\frac{\hat{w}}{\bar{x}}\right)^4 (-2\bar{x}\bar{y} s_{\bar{x}\bar{y}})$
$\frac{\bar{x}}{1 + \bar{x}}$	$\frac{s_{\bar{x}}^2}{(1 + \bar{x})^4}$	
$\bar{x}\bar{y}$	$(\bar{x}\bar{y})^2 \left(\frac{s_{\bar{x}}^2}{\bar{x}^2} + \frac{s_{\bar{y}}^2}{\bar{y}^2}\right)$	$(\bar{x}\bar{y})^2 \left(2\frac{s_{\bar{x}\bar{y}}}{\bar{x}\bar{y}}\right)$
$\bar{x}^2$	$4\bar{x}^2 s_{\bar{x}}^2$	
$\sqrt{\bar{x}}$	$\frac{1}{4}\frac{s_{\bar{x}}^2}{\bar{x}}$	
$\ln \bar{x}$	$\frac{s_{\bar{x}}^2}{\bar{x}^2}$	
$k\bar{x}^a\bar{y}^b$	$\hat{w}^2 \left(a^2 \frac{s_{\bar{x}}^2}{\bar{x}^2} + b^2 \frac{s_{\bar{y}}^2}{\bar{y}^2}\right)$	$(\hat{w})^2 \left(2ab \frac{s_{\bar{x}\bar{y}}}{\bar{x}\bar{y}}\right)$
$\dagger e^{a\bar{x}}$	$e^{2a\bar{x}} s_{\bar{x}}^2$	
$\frac{\sin\left(\frac{\bar{x} + \bar{y}}{2}\right)}{\sin\frac{\bar{x}}{2}}$	$\frac{1}{4}\left\{\frac{\cos^2\left(\frac{\bar{x} + \bar{y}}{2}\right)}{\sin^2\frac{\bar{x}}{2}} s_{\bar{u}}^2 + \frac{\sin^2\frac{\bar{y}}{2}}{\sin^4\frac{\bar{x}}{2}} s_{\bar{x}}^2\right\}$ ($s_{\bar{x}}$ and $s_{\bar{y}}$ in radians)	$-\frac{\sin\frac{\bar{y}}{2}\cos\left(\frac{\bar{x} + \bar{y}}{2}\right)}{2\sin^3\frac{\bar{x}}{2}} s_{\bar{x}\bar{y}}$
$100\frac{s_{\bar{x}}}{\bar{x}}$ (= coefficient of variation in percent)	$\frac{\hat{w}^2}{2(n-1)}$ (not directly derived from the formulas) ††	

* It is assumed that the value of $\hat{w}$ is finite and real, e.g., $\bar{y} \neq 0$ for ratios with $\bar{y}$ as denominator, $\bar{x} > 0$ for $\sqrt{\bar{x}}$ and $\ln \bar{x}$.

** Weighted mean as a special case of $\bar{x} + B\bar{y}$, with $\sigma_{\bar{x}}$ and $\sigma_{\bar{y}}$ considered known.

† Distribution of $\hat{w}$ is highly skewed and normal approximation could be seriously in error for small n.

†† See, for example, Statistical Theory with Engineering Applications, p. 301, by A. Hald (John Wiley & Sons, New York, N.Y., 1952).

gestions are made on the use of the standard deviation calculated for $\hat{u}$ when the standard deviations of x and y are not known. Readers may wish to go directly to paragraph 3.5 for a summary of the conclusions.

3.2. For x, y independently distributed and arbitrary $F(x, y)$, the first correction terms to (2.6) are

$$\left[\frac{\partial F}{\partial X}\right]\left[\frac{\partial^2 F}{\partial X^2}\right]\gamma_x\sigma_x^3+\left[\frac{\partial F}{\partial Y}\right]\left[\frac{\partial^2 F}{\partial Y^2}\right]\gamma_y\sigma_y^3. \tag{3.1}$$

where γ is a measure of skewness of the distribution.[5] Therefore these terms equal zero for x, y symmetrically distributed, a condition satisfied by assumption B_1.

The next order of correction terms involve σ_x^4, σ_y^4 and $\sigma_x^2\sigma_y^2$ and are usually negligible compared to terms in (2.6). These terms are

$$\frac{1}{3}\left\{\left[\frac{\partial F}{\partial X}\right]\left[\frac{\partial^3 F}{\partial X^3}\right]\Gamma_x\sigma_x^4+\left[\frac{\partial F}{\partial Y}\right]\left[\frac{\partial^3 F}{\partial Y^3}\right]\Gamma_y\sigma_y^4\right\}$$

$$+\frac{1}{4}\left\{\left[\frac{\partial^2 F}{\partial X^2}\right]^2(\Gamma_x-1)\sigma_x^4+\left[\frac{\partial^2 F}{\partial Y^2}\right]^2(\Gamma_y-1)\sigma_y^4\right\}$$

$$+\left\{\left[\frac{\partial F}{\partial X}\right]\left[\frac{\partial^3 F}{\partial X\partial Y^2}\right]+\left[\frac{\partial^2 F}{\partial X\partial Y}\right]^2+\left[\frac{\partial^3 F}{\partial X^2\partial Y}\right]\left[\frac{\partial F}{\partial Y}\right]\right\}\sigma_x^2\sigma_y^2. \tag{3.2}$$

For functions involving powers of x and y less than three, some of the partial derivatives also vanish. For example, if $W=XY$, the only nonzero term of this order is $\sigma_x^2\sigma_y^2$, or

$$\text{Var}\,(w)=M_y^2\sigma_x^2+M_x^2\sigma_y^2+\sigma_x^2\sigma_y^2.$$

The contribution of $\sigma_x^2\sigma_y^2$ is less than 1 in 200 if M/σ is larger than ten.

For functional forms such as quotients, roots, and logarithms, the accuracy is usually adequate since powers of the means of the variables appear in the denominators of the partial derivatives.

For the exponential function $W=e^X$, the variance of w as given by (2.6) is

$$\text{Var}\,(w)\doteq e^{2M}\sigma^2,$$

whereas the exact formula[6] for the variance of w, when x is normally distributed, is

$$\text{Var}\,(w)=e^{\sigma^2}e^{2M}(e^{\sigma^2}-1)$$

$$=e^{\sigma^2}e^{2M}\left(\sigma^2+\frac{\sigma^4}{2!}+\frac{\sigma^6}{3!}+\cdots\right)$$

$$=e^{2M}\sigma^2\left\{e^{\sigma^2}\left(1+\frac{\sigma^2}{2!}+\frac{\sigma^4}{3!}+\cdots\right)\right\}.$$

Here the variance of w as given by (2.6) underestimates the true variance by the factor given in the brackets, and the approximation could be seriously in error. (Note, however, the "exact" formula is correct only if x is exactly normally distributed. If x is only approximately normally distributed, then both formulas are approximations.)

For specific functions, formulas (3.1) and (3.2) given in Tukey's report can be used to check on the adequacy of the approximation. We quote Tukey's conclusion in this respect:

"The most important conclusion is that the classical propagation formula is much better than seems to be usually realized. Examples indicate that it is quite likely to suffice for most work."

3.3 Next we look into the adequacy of the normal approximation. For this purpose we will define the first four central moments of the distribution of w as follows:

$$\begin{aligned}&\text{mean}\,(w-M_w)=0\\&\text{mean}\,(w-M_w)^2=\sigma^2\\&\text{mean}\,(w-M_w)^3=\gamma\sigma^3\\&\text{mean}\,(w-M_w)^4=\Gamma\sigma^4.\end{aligned} \tag{3.3}$$

If w is normally distributed, $\gamma=0$, and $\Gamma=3$. Following Tukey, we shall define

$$\begin{aligned}\text{skewness}&=\gamma\sigma^3,\ \text{and}\\ \text{elongation}&=\Gamma\sigma^4-3\sigma^4;\end{aligned}$$

then both skewness and elongation are equal to zero when w is normally distributed.

If x and y are normally distributed as assumed under B_1, then in general $w=F(x, y)$ is not normally distributed unless the function form is linear. By a procedure similar to that used in the last section, the coefficients of skewness β_1 and excess β_2 of w can be calculated where:

$$\beta_1=\frac{[\text{skewness}\ w]^2}{[\text{var}\ w]^3}$$

$$\beta_2=\frac{\text{elongation}\ w}{[\text{var}\ w]^2}+3.$$

If β_1 is close to zero and β_2 is close to 3, the normal approximation may be considered as adequate.

The terms up to order σ^4 in the propagation of skewness for $w=F(x, y)$, with x, y independent, are

$$\text{skewness}\ w\doteq\left[\frac{\partial F}{\partial X}\right]^3\gamma_x\sigma_x^3+\left[\frac{\partial F}{\partial Y}\right]^3\gamma_y\sigma_y^3$$

$$+\frac{3}{2}\left[\frac{\partial F}{\partial X}\right]^2\left[\frac{\partial^2 F}{\partial X^2}\right](\Gamma_x-1)\sigma_x^4$$

$$+\frac{3}{2}\left[\frac{\partial F}{\partial Y}\right]^2\left[\frac{\partial^2 F}{\partial Y^2}\right](\Gamma_y-1)\sigma_y^4$$

$$+6\left[\frac{\partial F}{\partial X}\right]\left[\frac{\partial F}{\partial Y}\right]\left[\frac{\partial^2 F}{\partial X\partial Y}\right]\sigma_x^2\sigma_y^2. \tag{3.4}$$

[5] For definition of γ and Γ see (3.3).

[6] See, for example, The Lognormal Distribution, p. 8, by J. Aitchison and J. A. C. Brown, Cambridge University Press, 1957.

For x, y normally distributed, only terms of order σ^4 remain. If we take $w=xy$ again as an example, then

$$\text{skewness } w \doteq 6M_xM_y\sigma_x^2\sigma_y^2$$

$$\beta_1 \doteq \frac{36M_x^2M_y^2\sigma_x^4\sigma_y^4}{[M_y^2\sigma_x^2+M_x^2\sigma_y^2+\sigma_x^2\sigma_y^2]^3}.$$

Neglecting $\sigma_x^2\sigma_y^2$ in the brackets in the denominator, and taking $M/\sigma=10$, β_1 is computed to be 0.045. Hence, for $\hat{w}=\bar{x}\,\bar{y}$, where $\bar{x}$ and $\bar{y}$ are averages of four, the coefficient of skewness is reduced by a factor of four or equals 0.011 approximately.

Similarly, terms up to order σ^4 for the elongation of $w=f(x, y)$, with x, y independent, are

$$\text{elongation } w \doteq \left[\frac{\partial F}{\partial X}\right]^4(\Gamma_x-3)\sigma_x^4+\left[\frac{\partial F}{\partial Y}\right]^4(\Gamma_y-3)\sigma_y^4 \tag{3.5}$$

which is zero for x, y normal.

Hence $\beta_2=\dfrac{\text{elongation } w}{(\text{Variance } w)^2}+3=3$, and no correction for elongation is necessary here.

If we look up a table[7] of percentage points of distribution of the standardized variate $\dfrac{\hat{w}-M_w}{\sigma_{\hat{w}}}$ with given β_1 and β_2, we note that the changes of values are rather sensitive to β_1 and much less so to β_2. Thus the coefficient of elongation is usually not as much a source of worry in the normal approximation as is the coefficient of skewness.

Formulas (3.4) and (3.5) and the table of percentage points allow us to check how good the normal approximation is for a given number of measurements in the variables x and y. Table 2 gives some examples of results of such calculations.

3.4 The third approximation concerns the use of the sample variance s^2 as an estimate of the population variance σ^2. If we know the precision of the processes for the measurements of x and y, i.e., we know σ_x and σ_y, σ_w can be computed from (2.7) and a confidence interval about w can be constructed with the desired confidence coefficient $1-\alpha$ by using the table of the normal probability integral. If σ_x and σ_y are not known, then even if $\hat{\sigma}_{\hat{w}}$ can be computed from (2.9), the constants to be used for constructing a confidence interval with confidence coefficient $1-\alpha$ will be different from those for known σ.

To offer some guideline to the solution of this problem, we again assume measurements on x and y to be independently and normally distributed. If the number of measurements is large (a rule of thumb could be $n>30$), then (2.7) can be used assuming σ_x^2, σ_y^2, and σ_{xy} are known.

Of course one can always compute the half-widths of the respective 100 $(1-\alpha)$ percent confidence intervals for M_x and for M_y by the use of the Student's t statistic, and use (2.12) to get the half-width of the interval for M_w, i.e., set

$$\Delta x=t_{\left(1-\frac{\alpha}{2}\right),\,n-1}\frac{s_x}{\sqrt{n}} \quad \text{and,} \quad \Delta y=t_{\left(1-\frac{\alpha}{2}\right),\,k-1}\frac{s_y}{\sqrt{k}}.$$

and use (2.12) to get Δw. Then the interval $w\pm\Delta w$ is a confidence interval for M_w for a confidence coefficient of at least $(1-\alpha)$. This procedure, however, may be criticized on the ground of gross inefficiency in using the data.

We may write (2.9) as

$$\hat{\text{var}}\,(\hat{w})=\lambda_1s_x^2+\lambda_2s_y^2$$

where $\lambda_1=\dfrac{1}{n}\left[\dfrac{\partial F}{\partial X}\right]^2$ and $\lambda_2=\dfrac{1}{k}\left[\dfrac{\partial F}{\partial Y}\right]^2$ are two constants. For given degrees of freedom for s_x, $n-1$, and s_y, $k-1$, and given ratios of $\dfrac{\lambda_1s_x^2}{\lambda_1s_x^2+\lambda_2s_y^2}$, values of a "$v$" statistic have been tabulated[8] for confidence coefficients of 0.99, 0.98, 0.95, and 0.90. The interval

$$\hat{w}\pm v\sqrt{\lambda_1s_1^2+\lambda_2s_2^2} \tag{3.6}$$

is a confidence interval with confidence coefficient $1-\alpha$.

These tables, however, do not contain values for "v" for n and k less than 10, 10, 8, and 6 for the respective confidence coefficients, and hence cannot be used for smaller samples. In addition, they are useful only for two independent variables x and y.

Alternatively Welch [1947] has proposed the use of "effective degrees of freedom" for the estimated variance of $\hat{w}$ of the form

$$\hat{\text{var}}\,(\hat{w})=\Sigma\lambda_is_i^2.$$

The effective degree of freedom f is computed from

$$f=\frac{(\Sigma\lambda_is_i^2)^2}{\Sigma(\lambda_i^2s_i^4/f_i)} \tag{3.7}$$

where f_i is the degrees of freedom for[9] s_i^2.

In general f will be fractional. The t value with f degrees of freedom can be found or interpolated from the t table and the confidence interval computed as

$$\hat{w}\pm t_{\left(1-\frac{\alpha}{2}\right),\,f}\hat{\sigma}_{\hat{w}}.$$

[7] See Table 42, Biometrika Tables for Statisticians, Vol. 1, edited by E. S. Pearson and H. O. Hartley, The University Press, 1958. Also, pp. 79–84.

[8] See Table 11, Biometrika Tables for Statisticians, Vol. 1; also Further critical values for the two-means problem, W. H. Trickett, B. L. Welch, and G. S. James, Biometrika 43, 1956, pp. 204–5.

[9] If s_i^2 is computed from n_i measurements, the degrees of freedom is n_i-1.

TABLE 2. *Departures from normal approximations*

x, y independently distributed, with $\gamma = 0$, $\Gamma = 3$, and $(M/\sigma) = 10$.

$\hat{u}$	Skewness from (3.4)	β_1 computed	Percentage point of $\frac{\hat{u} - M_u}{\hat{\sigma}_{\hat{u}}}$	
			Lower 2.5%	Upper 2.5%
$Ax + By$	0	0	−1.96	+1.96 †
xy	$6M_xM_y\frac{\sigma_x^2\sigma_y^2}{n^2}$	$\frac{4.5}{100n}$		
$n = 1$		0.045	−1.84	+2.06
$n = 4$		.011	−1.91	+2.01
$n = 10$		.0045	−1.93	+1.99
x^2	$24M^2\frac{\sigma^4}{n^2}$	$\frac{9}{100n}$		
$n = 10$		0.009	−1.90	+2.00
x^3	$162M^6\frac{\sigma^4}{n^2}$	$\frac{36}{100n}$		
$n = 4$		0.09	−1.80	+2.09
x^4	$6\left(\frac{M_x}{M_y}\right)^3\left(\frac{\sigma_y^4}{M_y^4} + \frac{\sigma_x^2\sigma_y^2}{M_x^2M_y^2}\right)$	$\frac{18}{100n}$		
$n = 10$		.018	−1.89	+2.03
$\ln x$ **	$\frac{3}{M^4}\cdot\frac{\sigma^4}{n^2}$	$\frac{9}{100n}$		
$n = 10$		0.009	−1.90	+2.00
e^x	$3e^{3M}\frac{\sigma^4}{n^2}$	$\frac{9\sigma^2}{n}$	Depends on σ and n (both skewness and β_1 underestimated for $\sigma/\sqrt{n} > 0.2$).	

* $\gamma > 0$.

** $x > 0$.

† Exact when x and y are normally distributed.

The approximate confidence intervals computed by the use of effective degrees of freedom were found to check the exact confidence intervals given by (3.6) very well over the range of the latter.

3.5 In summary, the following may be concluded for practical purposes:

(1) Terms of order higher than σ^2 in the propagation of error formulas for variance, (2.6) and (2.7), can be neglected if (a) the standard deviations are small in comparison to their respective means, and (b) the second and higher order partial derivatives evaluated at the means do not give rise to abnormally large numbers. This is usually true in the field of physical science, since errors of measurements are usually of the order of 1 part in 1000, or parts per million: furthermore, the functional forms used are usually the well-behaved ones.

(2) The normal approximation will be adequate for large n, or if, in addition to (a) and (b) above, (c) the individual variables can be assumed to be normally distributed. For particular functions, the approximate values of the coefficients of skewness and elongation may be calculated and Pearson's table can be used to check the adequacy of the approximation.

(3) For the case where the standard deviations of the individual variables are unknown, and are estimated from the data, confidence intervals for the estimate $\hat{u}$ can be constructed either by the use of tabulated values of the "v" statistic or by the use of effective degrees of freedom. These confidence intervals can be considered as a form of "precision limits" in the sense that if one makes the same sets of measurements a large number of times under the same conditions, and constructs the confidence intervals each time by the same procedure, then a large proportion of the intervals so constructed, $100(1-\alpha)$ percent, will bracket the mean of all these sets of measurements. When only one set of measurements will be made, the probability is $1-\alpha$ that this interval will bracket the mean.

4. Reporting of Results

4.1. Suppose a set of measurement data is available, and, by using the appropriate propagation of error formulas, the following are obtained for the quantity of interest, w_0:

(1) The estimate of w_0, $\hat{w}$, based on n values of x,y, etc.;

(2) the estimated standard error of $\hat{w}$, $\hat{\sigma}_w$, and associated degress of freedom f;

(3) limits to the systematic error in w, Δw.

The estimated standard error of $\hat{w}$ gives a measure of precision of the experimental results, or a measure of scatter of the values of w from the average value of M_w for repeated performance of the particular experiment. But this measure of precision does not indicate at all how close this average value is to the value w_0 intended to be measured. The estimation of limits to the systematic error is an essential part of an experiment and need not be discussed here [Youden, 1961]. One may remark generally that systematic errors usually do not pose a serious problem when the "imprecision" is large, since these systematic errors are, so to speak, "swallowed up" by the random errors. The systematic errors, however, play an important role when the precision is excellent and is of about the same order of magnitude as the systematic error. In that case, it is essential that the systematic error, or errors, be reported separately from the imprecision part of the reported value, as measured by the standard error, or the confidence intervals, computed.

In scientific literature, it is not uncommon to come across expressions of results in the form of $M \pm e$, where "M" is an average of some kind and "e" represents the uncertainty of "M" in some vague sense. This type of reporting proves to be most frustrating from the reader's point of view. From the context alone the reader cannot possibly infer whether "e" represents probable error, 3-sigma limits, systematic error, or some combination of random and systematic errors. As a consequence, the quality of the results, and the validity of inference drawn from these results, are to a large extent left to the judgment and guesswork of the reader. Hence, the writer owes to himself, and to his reader, to specify clearly the meaning of "e" as he uses it. In particular, the number of measurements from which the measure of random error was computed and the manner in which the systematic error was estimated are both essential elements of the reported value and need to be included.

A footnote explaining the role of "e" is often very helpful. Several examples are given below:

"In the expression of the form $M \pm e$, M is the average and e is the standard error of M based on n measurements (or based on ν degrees of freedom)."

"The indicated uncertainty limits for M are overall limits of error based on 95 percent confidence limits for the mean [illegible] and on allowances for effects of known sources of possible systematic error [illegible]."

"The uncertainty given represents 3-sigma limits based on the current accepted value of the standard deviation, known sources of systematic errors being negligible."

Chapter 23 of Natrella [1963] "Expressions of the Uncertainties of Final Results" gives a thorough discussion on this topic, and is an excellent reference for all physical scientists who have occasion to report numerical results of their experiments.

5. References

Birge, Raymond T., The propagation of errors, The American Physics Teacher **7**, No. 6 (Dec. 1939).

Craig, C. C., On frequency function of XY, Annals of Math. Stat. **7**, 1–15 (1937).

Cramér, Harald, Mathematical Methods of Statistics (Princeton University Press, 1946).

Curtiss, J. H., On transformations used in the analysis of variance, Annals of Math. Stat. **14**, 107–122 (1943).

Doob, J. L., The limiting distribution of certain statistics, Annals of Math. Stat. **6**, 160–170 (1935).

Eisenhart, Churchill, Realistic evaluation of the precision and accuracy of instrument calibration systems, J. Res. NBS **67C** (Engr. and Instr.) No. 2, 161–187 (1963).

Fieller, E. C., The distribution of the index in a normal bivariate population, Biometrika **24**, 428–440 (1932).

Goodman, L. A., The variance of the product of K random variables, J. Am. Stat. Assoc. **57**, 54–60 (1962).

Hsu, P. L., The limiting distribution of functions of sample means and application to testing hypotheses, pp. 359–402, Berkeley Symposium on Mathematical Statistics and Probability, Univ. of California Press (1949).

Mandel, John, The Statistical Analysis of Experimental Data (John Wiley & Sons, Inc., New York, N.Y., 1964).

Natrella, Mary Gibbons, Experimental Statistics, Handbook 91, National Bureau of Standards, 1963.

Tukey, John W., The propagation of errors, fluctuations and tolerances, Unpublished Technical Reports No. 10, 11, and 12, Princeton University.

Welch, B. L., The generalization of 'Student's' problem when several different population variances are involved, Biometrika **34**, 28–35 (1947).

Youden, W. J., Systematic errors in physical constants, Physics Today **14** (Sept. 1961).

(Paper 70C4–237)

RANDOMIZATION IN FACTORIAL AND OTHER EXPERIMENTS

E. Bright Wilson, Jr.

4.10. Randomization in Factorial and Other Experiments

The principle of randomization has important applications to factorial design. Besides the variables forming the factors which are consciously varied from observation to observation, there are always other variables which vary either in an unknown way or with unknown effects. The less the influence of these variables, the more precise the experiment, but in no experiment can their effects be completely eliminated. Wherever possible these other variables should be *randomized.*

For example, in repetitions of measurements with complex physical apparatus, it is seldom possible to carry out experiments all at the same time. *Time* therefore is another variable which is not held constant during the various comparisons, and it is well known that many disturbing variables such as temperature, line voltage, state of chemical decomposition, quantity of living matter, etc., can change with time. In other cases sets of experiments can be carried out at the same time but then not in the same *place,* or with the same equipment, or on the same patients. The effects of any of these variables may be confused with those of the factors under test.

To reduce this danger, the temporal order of the experiments, the apparatus, the patients, or the position, etc., used with each combination of factors should be chosen by a truly random process such as the application of a table of random numbers. (See Sec. 10.3.)

Two Examples. In an important set of chemical analyses, it was standard practice to follow each analysis by a duplicate run. The agreement between the pairs was good, and so the results were accepted with confidence. However, their importance was sufficient so that samples were sent for analysis to an independent laboratory, with the result that wide discrepancies were found between the analyses from the two places. Investigation showed that a zinc reductor, through which all the samples were passed in turn, gradually lost its effectiveness because of the presence of certain other elements in the samples. The effect from one analysis to the next was small, so that the pairs checked well, but by the end of a day the absolute values were much in error. If the temporal order in which the analyses were made had been randomized, many of the pairs would have been widely separated in time and the agreement between pairs would not have been regarded as adequate.

This point is so important that another example may not be superfluous. In an industrial laboratory, experiments were performed to

determine the effect of the length of time of pressing in the mold on the strength of a plastic part. Hot plastic was introduced in the mold, pressed for 10 seconds, and removed. Another batch was then introduced into the same mold, pressed for 20 seconds, and so on, the time increasing with each batch. Afterward the strength of each piece was measured and plotted against the duration of the pressure. Figure 4.1

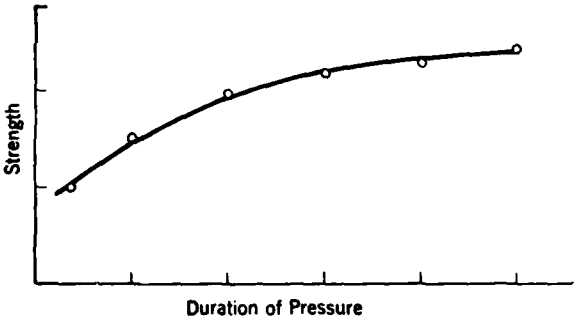


FIG. 4.1. Results of an experiment supposed to demonstrate that the strength of a molded plastic part depended on duration of pressure. See Fig. 4.2.

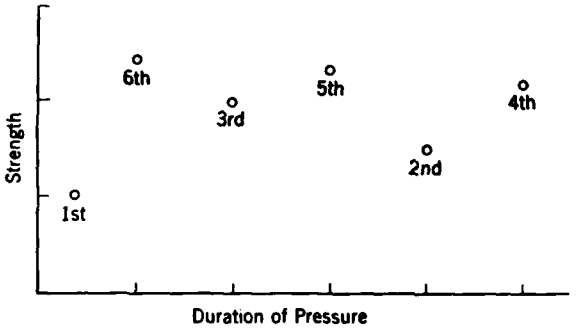


FIG. 4.2. Repeat of experiment shown in Fig. 4.1 except that order in which experiments were carried out was randomized. Order is shown. Results depended on order, not on duration.

shows the resulting curve, which was taken to indicate a strong dependence of strength on duration. However, the research supervisor criticized the experiment because the order of the experiments had not been randomized, and so it was repeated. The results are shown in Fig. 4.2, which also notes the order in which the measurements were taken. Obviously, it was the *order* and not the *duration* which was the controlling variable; the first conclusion was quite erroneous. The origin of the trouble was easily traced after its presence was made known; the mold got warmer and warmer as successive batches of hot plastic were pressed in it.

Other Variables. Time is not the only variable which should be randomized. In agricultural experiments the exact position of each test variety in a plot is randomly selected. If different materials or specimens are used, they should be randomly selected. For example, in testing explosives steel plates were used in the gauges. These were rolled from one ingot to promote uniformity, but the whole batch of plates was shuffled thoroughly before use to randomize out any bias due to variations in strength.

Randomization converts effects of randomized variables into unbiased error. The importance of randomization in connection with the mathematical methods of estimating error and the significance of differences will be discussed in Sec. 8.6.

If there is strong reason to believe that a certain variable will influence the results of an experiment and if it can be controlled, it should be included in the design as one of the factors; but if it is only suspected that it might have some influence or if it cannot be controlled, then randomization is the safe course. Sometimes both techniques are combined; crop varieties may be planted in parts of plots, the parts being randomly selected, but the separate plots, being farther apart and therefore probably more divergent in fertility, are considered as different values of the factor position. Similarly groups of observations taken near to one another in time may be randomized as to order, but repetitions of the groups (with separately randomized orders) which are more widely spaced in time may be regarded as testing the effect of the factor time.

When any element of an experiment has been randomized, it is important to keep good records of the original situation. Thus it was important to know the order in which the plastic-molding experiments of Fig. 4.2 were carried out because afterward they could be reordered to show that a trend existed. This is often the case in more complicated factorial experiments, as will be discussed in Sec. 4.12.

Balanced Designs. From time to time objections have been raised against the use of randomization. It has been argued that it is better to use consciously planned patterns devised to minimize errors due to the extra variables. For example, in a simple case of an experiment in which a variety A is compared with B with several replications, in a linear physical layout or in a time sequence or with a similar linear variation of some other variable, a possible randomized arrangement would be

$$A\ B\ B\ A\ A\ A\ B\ A\ B\ B\ A\ B$$

But it could happen that the random choice produced the pattern

$$A\ A\ A\ A\ A\ A\ B\ B\ B\ B\ B\ B$$

If it was suspected that a trend existed along the line (as a fertility

gradient in a field, or a trend with time), this pattern would be a poor one, yet it could have resulted from the random draw.

For this reason some authorities have recommended systematic arrangements, such as

$$A\,B\,B\,A\,A\,B\,B\,A\,A\,B\,B\,A$$

in which such a gradient would tend to be balanced out. However, there are very strong arguments against these "balanced" arrangements. For example, although a gradient is taken care of by the above pattern, a periodicity with a period coinciding with that apparent in the pattern could cause serious bias. Such periods are easily introduced in certain types of experiment.

In many cases both schools can be satisfied. Instead of randomizing without restriction, certain restrictions may be introduced. Thus it may be fixed that every adjacent pair contains one A and one B but the order in each pair is chosen randomly. This really amounts to introducing the gross position along the sequence as an additional factor and randomizing only within each "position." Perhaps the best rule in many cases is to select by lot one from among all the possible patterns which as far as available foresight is concerned are equally good. These should of course not be biased; *i.e.*, for every pattern included there should also be included any which result from a permutation of treatments throughout, *e.g.*, exchanging A and B above. In applying the mathematical theorems of Sec. 8.11 to these cases, certain rules need to be observed which may limit the freedom with which patterns are discarded.

From Wilson, E. B., An Introduction to Scientific Research, Section 4.10, pp. 54-57 (McGraw-Hill Book Co., New York, 1952).

Some Remarks on Wild Observations*

WILLIAM H. KRUSKAL**

The University of Chicago

Editor's Note: At the 1959 meetings of the American Statistical Association held in Washington D. C., Messrs. F. J. Anscombe and C. Daniel presented papers on the detection and rejection of 'outliers', that is, observations thought to be maverick or unusual. These papers and their discussion will appear in the next issue of *Technometrics*. The following comments of Dr. Kruskal are another indication of the present interest of statisticians in this important problem.

The purpose of these remarks is to set down some non-technical thoughts on apparently wild or outlying observations. These thoughts are by no means novel, but do not seem to have been gathered in one convenient place.

1. Whatever use is or is not made of apparently wild observations in a statistical analysis, it is very important to say something about such observations in any but the most summary report. At least a statement of how many observations were excluded from the formal analysis, and why, should be given. It is much better to state their values and to do alternative analyses using all or some of them.

2. However, it is a dangerous oversimplification to discuss apparently wild observations in terms of inclusion in, or exclusion from, a more or less conventional formal analysis. An apparently wild (or otherwise anomalous) observation is a signal that says: "Here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study." Examples of such serendipity have been frequently discussed—one of the most popular is Fleming's recognition of the virtue of penicillium.

3. Suppose that an apparently wild observation is *really known* to have come from an anomalous (and perhaps infrequent) causal pattern. Should we include or exclude it in our formal statistics? Should we perhaps change the structure of our formal statistics?

Much depends on what we are after and the nature of our material. For example, suppose that the observations are five determinations of the percent of chemical A in a mixture, and that one of the observations is badly out of

* This work was sponsored by the Army, Navy and Air Force through the Joint Services Advisory Committee for Research Groups in Applied Mathematics and Statistics by Contract No. N6ori-02035. Reproduction in whole or in part is permitted for any purpose of the United States Government.

** With generous suggestions from L. J. Savage, H. V. Roberts, K. A. Brownlee, and F. Mosteller.

line. A check of equipment shows that the out of line observation stemmed from an equipment miscalibration that was present only for the one observation.

If the magnitude of the miscalibration is known, we can probably correct for it; but suppose it is not known? If the goal of the experiment is only that of estimating the per cent of A in the mixture, it would be very natural simply to omit the wild observation. If the goal of the experiment is mainly, or even partly, that of investigating the *method* of measuring the per cent of A (say in anticipation of setting up a routine procedure to be based on one measurement per batch), then it may be very important to keep the wild observation in. Clearly, in this latter instance, the wild observation tells us something about the frequency and magnitude of serious errors in the method. The kind of lesson mentioned in 2 above often refers to methods of sampling, measurement, and data reduction, instead of to the underlying physical phenomenon.

The mode of formal analysis, with a known anomalous observation kept in, should often be different from a traditional means-and-standard deviations analysis, and it might well be divided into several parts. In the above very simple example, we might come out with at least two summaries: (1) the mean of the four good observations, perhaps with a $\pm$ attached, as an estimate of the per cent of A in the particular batch of mixture at hand, and (2) a statement that serious calibration shifts are not unlikely and should be investigated further. In other situations, nonparametric methods might be useful. In still others, analyses that suppose the observations come from a mixture of two populations may be appropriate.

The sort of distinction mentioned above has arisen in connection with military equipment. Suppose that 50 bombs are dropped at a target, that a few go wildly astray, that the fins of these wild bombs are observed to have come loose in flight, and that their wildness is unquestionably the result of loose fins. If we are concerned with the accuracy of *the whole* bombing system, we certainly should not forget these wild bombs. But if our interest is in the accuracy of the bombsight, the wild bombs are irrelevant.

4. It may be useful to classify different degrees of knowledge about an apparently wild observation in the following way:

a. We may know, even *before* an observation, that it is likely to be wild, or at any rate that it will be the consequence of a variant causal pattern. For example, we may see the bomb's fins tear loose before it has fallen very far from the plane. Or we may know that a delicate measuring instrument has been jarred during its use.

b. We may be able to know, *after* an observation is observed to be apparently outlying, that it was the result of a variant causal pattern. For example, we may check a laboratory notebook and see that some procedure was poorly carried out, or we may ask the bombardier whether he remembers a particular bomb's wobbling badly in flight. The great danger here, of course, is that it is easy after the fact to bias one's memory or approach, knowing that the observation seemed wild. In complex measurement situations we may often find something a bit out of line for almost any observation.

c. There may be *no evidence* of a variant causal pattern aside from the observa-

tions themselves. This is perhaps the most difficult case, and the one that has given rise to various rules of thumb for rejecting observations.

Like most empirical classifications, this one is not perfectly sharp. Some cases, for example, may lie between b and c. Nevertheless, I feel that it is a useful trichotomy.

5. In case c above, I know of no satisfactory approaches. The classical approach is to create a test statistic, chosen so as to be sensitive to the kind of wildness envisaged, to generate its distribution under some sort of hypothesis of non-wildness, and then to 'reject' (or treat differently) an observation if the test statistic for it comes out improbably large under the hypothesis of nonwildness. A more detailed approach that has sometimes been used is to suppose that wildness is a consequence of some definite kind of statistical structure—usually a mixture of normal distributions—and to try to find a mode of analysis well articulated with this structure.

My own practice in this sort of situation is to carry out an analysis both with and without the suspect observations. If the broad conclusions of the two analyses are quite different, I should view any conclusions from the experiment with very great caution.

6. The following references form a selected brief list that can, I hope, lead the interested reader to most of the relevant literature.

References

1. C. I. Bliss, W. G. Cochran, and J. W. Tukey, "A rejection criterion based upon the range," *Biometrika*, 43 (1956), 418–22.
2. W. J. Dixon, "Analysis of extreme values," *Ann. Math. Stat.*, 21 (1950), 488–506.
3. W. J. Dixon, "Processing data for outliers," *Biometrics*, 9 (1953), 74–89.
4. Frank E. Grubbs, "Sample criteria for testing outlying observations," *Ann. Math. Stat.*, 21 (1950), 27–58.
5. E. P. King, "On some procedures for the rejection of suspected data," *Jour. Amer. Stat. Assoc.*, 48 (1953), 531–3.
6. Julius Lieblein, "Properties of certain statistics involving the closest pair in a sample of three observations," *Jour. of Research of the Nat. Bureau of Standards*, 48 (1952), 255–68.
7. E. S. Pearson and C. Chandra Sekar, "The efficiency of statistical tools and a criterion for the rejection of outlying observations," *Biometrika*, 28 (1936), 308–320.
8. Paul R. Rider, "Criteria for rejection of observations," *Washington University Studies, New Series. Science and Technology*, 8 (1933).

Reprinted from AMERICAN JOURNAL OF PHYSICS, Vol. 21, No. 7, 520–525, October, 1953
Printed in U. S. A.

Rejection of Outlying Observations

FRANK PROSCHAN*
National Bureau of Standards, Washington, D. C.
(Received November 24, 1952)

This paper makes available to the physicist two of the modern statistical tests for possible rejection of outlying observations. These two methods have been selected because they apply in a majority of the actually occurring situations and because they are so easy to use.

A PERENNIAL problem vexing the experimenter is that of rejection of suspected data. For one hundred years attempts at the solution of this problem have been advanced, most of them to be themselves rejected as suspect. Fortunately, modern statistical theory has proposed useful, reliable methods for objectively rejecting deviant values. However, the solution is far from complete at present.

This paper makes available to the physicist two of the modern statistical tests for possible rejection of outlying observations. These two methods have been selected because they apply in a majority of the actually occurring situations and because they are so easy to use.

THE PROBLEM

Here is a common problem facing experimenters. The typical scientist, X. Perry Menter, makes a number (say five) of repeated measurements of some unknown quantity. The smallest value (or the largest) is so far removed from the other four that he suspects that it may be in error. However, Perry has no specific knowledge that a mistake actually did occur. Let us assume that he has no previous data from which to estimate the precision of measurement. How can he decide *from the values themselves* whether the suspected value is in error or not?

The answer seems clear. He should consider the suspected value as in error when it seems too far from the other four values. But how can he judge when it is "too far from the other four values"?

A LOGICAL APPROACH

Here is a simple, logical, objective criterion. Suppose Perry could somehow make millions of sets of five observations each. Suppose, too, that he could guarantee that none of these observations had any mistakes. Call a typical set x_1, x_2, x_3, x_4, x_5, where the x's are arranged in order of size, so that $x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5$. Now a logical measure of the distance between the smallest value and the other four values is

$$r_{10} = \frac{x_2 - x_1}{x_5 - x_1},$$

i.e., the ratio of the interval between the suspected and adjacent value to the total range.

Now Perry records with what frequency, among his millions of sets of five values each, different values of r_{10} occur. He finds that a value of r_{10} larger than 0.780 occurs one time in one hundred. He then reasons this way:

"I have found that among sets of five observations each (*containing no mistakes*) a value of r_{10} larger than 0.780 is quite unlikely (occurs only once in one hundred). If now, in my future experiments I get a set of five observations for which r_{10} is larger than 0.780, I will conclude that my largest observation is in error."

CONFIDENCE IN THE TEST

This seems reasonable. But what confidence can Perry have in such a procedure? How often will he consider as mistaken a perfectly good observation? How often will he consider acceptable an incorrect observation?

Clearly, from the way in which he derived the test, he will classify a perfectly good smallest observation as mistaken once among one hundred sets of five each, on the average. But there is no general answer to the question of how often he will let pass a mistaken observation. This depends on how "mistaken" the mistaken observation is. If a very *large* error were made, his

* Now at Sylvania Electric Products, Inc., Hicksville, New York.

test would tend to reject the observation almost certainly. If a very small error were made, his test would tend to reject the observation with a small probability.

Figure 1 gives some idea of the performance of r_{10} in detecting mistaken observations. It is based on a sampling experiment in which samples of five from a normal population with mean μ and standard deviation σ were contaminated with values drawn from a normal population with mean $(\mu+\lambda\sigma)$ and standard deviation σ. The ordinate shows the percent discovery of contaminators (the proportion of the time the contaminating population provides an extreme value and the test discovers this value) while the abscissa shows λ, the magnitude of the shift (error) of the contaminator in standard deviations.

We said above that once in every 100 sets of values (on the average) Perry would consider as mistaken a perfectly good observation. If he were to reject this observation and then compute the mean and standard deviation of the remaining values, these would be biased estimates. In addition, when a good observation is rejected, any further statistical tests of significance will become less reliable. This is the price that he must pay for improving the data in the cases where a mistaken observation is removed.

MATHEMATICAL DERIVATION

Of course, 0.780, the value of r_{10} that is exceeded by chance 1 percent of the time (called the 1 percent level of significance of r_{10}), is not determined by actually making millions of sets of five observations each. Rather it may be calculated mathematically[1] with even greater accuracy than if millions of sets of five observations had been used. The basic assumption is that the repeated measurements would follow the normal distribution.

LARGER SAMPLE SIZES

For sample sizes larger than seven, slight modifications in the r_{10} statistic result in a more

[1] Dixon, Ann. Math. Stat. 22, No. 1, 68–70 (1951).

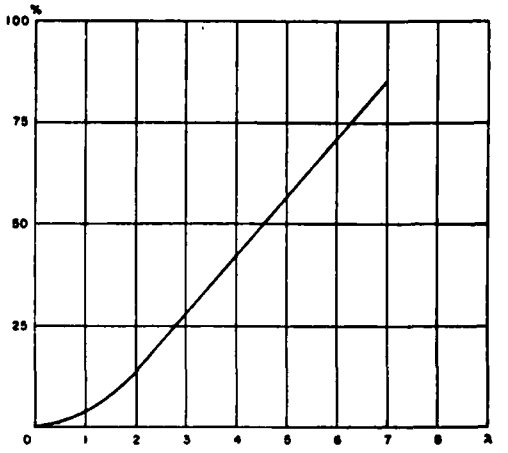


FIG. 1. Performance of r test. The ordinate shows the percent discovery of contaminators, while the abscissa shows λ, the magnitude of the shift (error) of the contaminator in standard deviations. From W. J. Dixon, Ann. Math. Stat. 21, No. 4, 493 (1950).

sensitive test: Thus for sample size $n=8$, 9, or 10,

$$r_{11}=\frac{x_2-x_1}{x_{n-1}-x_1}$$

is superior to r_{10}. Similarly for $n=11$, 12, or 13,

$$r_{21}=\frac{x_3-x_1}{x_{n-1}-x_1}$$

is superior. Finally for $n=14, 15, \cdots, 30$,

$$r_{22}=\frac{x_3-x_1}{x_{n-2}-x_1}$$

is best.

USE OF TABLE I

Let us now define r as the appropriate statistic among r_{10}, r_{11}, r_{21}, and r_{22} according to the sample size. Table I gives critical values of r for significance levels $\alpha=5$ percent and $\alpha=1$ percent, for sample sizes from $n=3$ to 30.

Thus for example for $n=8$ and $\alpha=5$ percent, the table gives a critical value for r (in this case r_{11}) of 0.554. This means that in 100 sets of 8 observations each, free of mistakes, five values of r_{11} will be larger than 0.554, on the average.

What if Perry suspects the acceptability of the largest observation in a set? In this case, he simply considers the observations as numbered in the reverse order and proceeds as before.

TABLE I. Testing for extreme observation (no past data).*

Statistic	Sample size n	Critical values $\alpha=5$ percent	$\alpha=1$ percent
$r_{10}=\frac{x_2-x_1}{x_n-x_1}$	3	0.941	0.988
	4	0.765	0.889
	5	0.642	0.780
	6	0.560	0.698
	7	0.507	0.637
$r_{11}=\frac{x_2-x_1}{x_{n-1}-x_1}$	8	0.554	0.683
	9	0.512	0.635
	10	0.477	0.597
$r_{21}=\frac{x_3-x_1}{x_{n-1}-x_1}$	11	0.576	0.679
	12	0.546	0.642
	13	0.521	0.615
$r_{22}=\frac{x_3-x_1}{x_{n-2}-x_1}$	14	0.546	0.641
	15	0.525	0.616
	16	0.507	0.595
	17	0.490	0.577
	18	0.475	0.561
	19	0.462	0.547
	20	0.450	0.535
	21	0.440	0.524
	22	0.430	0.514
	23	0.421	0.505
	24	0.413	0.497
	25	0.406	0.489
	26	0.399	0.486
	27	0.393	0.475
	28	0.387	0.469
	29	0.381	0.463
	30	0.376	0.457

* By permission from W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis* (McGraw-Hill Book Company, Inc., New York, 1951), p. 319.

Why are two significance levels given? The reason is that no one significance level is appropriate to all problems. For example, consider these two cases:

(a) Additional observations are not possible.
(b) Additional observations are possible.

In case (a) for many problems it might be appropriate to compute r and test it at the 1 percent level of significance. If the particular observed value of r is larger than the tabulated value for $\alpha=1$ percent, it might then be a good idea to exclude that observation.

In case (b), for many situations a reasonable procedure might be to test r at the 5 percent level. If the sample value of r is significant at the 5 percent level, one or more additional observations would be taken. If the observation originally suspected remained outlying, it would be tested again, using the combined set of observations. This time, however, the r test would be performed at the *1 percent* level of significance. If the outlier were significantly deviant at the 1 percent level, it would be rejected. It should be noted that among many sets tested in this way, the proportion of sets in which a perfectly good largest value will thus be rejected will be less than 1 percent. This is because the observation has a "second chance" before it is finally rejected.

SUMMARY

A set of n observations is made. No previous data are available from which to estimate the variability of a measurement. What is a rational procedure for testing whether the largest (or smallest) of the set is too deviant to be explained by the ordinary errors of measurement?

Rank the n observations in order of size from smallest to largest, if the smallest observation is suspected,

$$x_1 \leq x_2 \leq \cdots \leq x_n;$$

reverse the numbering system if the largest is suspected.

Next compute

$$r_{10}=\frac{x_2-x_1}{x_n-x_1} \quad \text{if } n=3 \text{ to } 7$$

$$r_{11}=\frac{x_2-x_1}{x_{n-1}-x_1} \quad \text{if } n=8 \text{ to } 10$$

$$r_{21}=\frac{x_3-x_1}{x_{n-1}-x_1} \quad \text{if } n=11 \text{ to } 13$$

$$r_{22}=\frac{x_3-x_1}{x_{n-2}-x_1} \quad \text{if } n=14 \text{ to } 30.$$

Table I may be used to determine how likely it is to get as large a value of r as actually obtained, simply by chance. A procedure that might be appropriate for many problems is as follows.

(*a*) *No additional observations possible.* In this case, compared the computed r with the value in Table I at the 1 percent level. If the computed value of r is larger than the tabulated value, exclude the deviant observation. Otherwise, do not.

(*b*) *Additional observations possible.* In this case, compare the computed r with the value of r at the 5 percent level. If the computed value of r is larger than the value, take one or more additional observations. Otherwise accept the suspected value without taking additional observations.

If, in the enlarged set (containing all the original and the additional observations), the previously suspected value remains outlying, compute r for the enlarged set. This time compare it with the value at the 1 percent level. If the computed value exceeds the table value, exclude the outlier; otherwise do not.

EXAMPLES

1. In a preliminary experiment, Silas N. Tist makes 5 determinations of the velocity of light in vacuum by a new method, obtaining 299 792, 299 780, 299 795, 299 786, 299 820, (km/sec). Si N. Tist suspects the last value, 299 820, as being mistaken since it is so much larger than the other values. Before going on with additional experimentation, Si wishes to decide whether 299 820 is mistaken or not. What shall he do?

Since no previous data are available from which to compute the precision of measurement by this new method, the r test is appropriate. The first step is to arrange the five values in order of size: 299 780, 299 786, 299 792, 299 795, 299 820. Then

$$r=r_{10}=\frac{299\,820-299\,795}{299\,820-299\,780}=\frac{25}{40}=0.625.$$

Since this is less than 0.780, the 0.01 point of r for $n=5$, Si N. Tist concludes that 299 820 is not mistaken.

2. Using the Atwood machine, Norris G. Neer makes determinations of g, the acceleration of gravity, in his college course in experimental physics. N. G. Neer's values are: 986, 964, 989, 1000, 987, 909, 999 (cm/sec^2). He suspects 909 as being inconsistent with the other values. Shall he accept it, or shall he experiment further?

He computes

$$r=r_{11}=\frac{x_2-x_1}{x_7-x_1}=\frac{964-909}{1000-909}=\frac{55}{91}=0.604.$$

This value lies between the 0.01 and the 0.05 points of r for $n=7$. Hence N. G. Neer makes an additional determination and gets a new value of 971.

Since 909 remains outlying in the enlarged set of eight, he computes r for this set of eight. Now $r(r_{11})$ is 0.611. Since it is smaller than the 1 percent level of r for $n=8$, N. G. Neer accepts 909 and uses all eight values.

ESTIMATE OF MEASUREMENT VARIABILITY AVAILABLE

In a great many laboratory situations, past data *are* available for estimating the uncertainty of a measurement. It is clear that where such information is available, it should be used in deciding whether an outlier is mistaken or not. This will make the decision more reliable than if only the one set containing the suspected value is used.

The u Test

The test ratio used now is

$$u=\frac{x_n-\bar{x}}{s_d}\quad\text{(If } x_n \text{ is the suspected value)}$$

or

$$u=\frac{\bar{x}-x_1}{s_d}\quad\text{(If } x_1 \text{ is the suspected value),}$$

where

$\bar{x}$ = mean of the set of observations,
s_d = standard deviation of an individual measurement, based on d degrees of freedom.

Calculating s_d

To determine s_d from a single set of measurements we would first calculate the sum of the squares of the deviations of the observations from their mean. Then we would divide by one less than the number of observations. This would give us an unbiased estimate of the variance s_d^2. Thus,

$$s_d^2=\left(\sum_{i=1}^{d}(x_i-\bar{x})^2\right)\Big/(d-1).$$

On the other hand, suppose a number of sets of

observations were available:

1	x_{11},	x_{12},	$\cdots$, x_{1n_1}	$\bar{x}_1$
2	x_{21},	x_{22},	$\cdots$, x_{2n_2}	$\bar{x}_2$
...	...	...	...	...
k	x_{k1},	x_{k2},	$\cdots$, x_{kn_k}	$\bar{x}_k$.

Now we could calculate s_d^2 from

$$s_d^2=\left(\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2+\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2+\cdots+\sum_{i=1}^{n_k}(x_{ki}-\bar{x}_k)^2\right)\Big/((n_1-1)+(n_2-1)+\cdots+(n_k-1)), \quad (1)$$

where d, the number of degrees of freedom for estimating the uncertainty of measurement, is $(n_1-1)+(n_2-1)+\cdots+(n_k-1)$.

An example: An example will make the whole procedure clear: Alec Tronick has just made three determinations of the frequency deviation sensitivity in megacycles/volt for a certain reflex klystron, obtaining the values: 1.13, 1.12, 1.02, with mean, $\bar{x}=1.09$.

He wishes to test whether "1.02" is unusually deviant. He has past data (Table II) to estimate the precision of this type of measurement.

Although he could calculate s_d^2 from Eq. (1), it is generally more convenient (especially with a computing machine available) to use the following Eq. (2).

$$s_d^2=\left[\sum_{i=1}^{n_1}x_{1i}^2-\frac{\left(\sum_{i=1}^{n_1}x_{1i}\right)^2}{n_1}+\sum_{i=1}^{n_2}x_{2i}^2-\frac{\left(\sum_{i=1}^{n_2}x_{2i}\right)^2}{n_2}+\cdots+\sum_{i=1}^{n_k}x_{ki}^2-\frac{\left(\sum_{i=1}^{n_k}x_{ki}\right)^2}{n_k}\right]\Big/(n_1-1)+(n_2-1)+\cdots+(n_k-1). \quad (2)$$

It can easily be shown that Eq. (2) is algebraically equivalent to Eq. (1). Thus, substituting values into Eq. (2) yields

$$s_d^2=[1.11^2+1.07^2+1.10^2-(1.11+1.07+1.10)^2/3+\cdots+1.06^2+1.01^2+1.08^2-(1.06+1.01+1.08)^2/3]/[(3-1)+\cdots+(3-1)]=0.0271/16=0.001694.$$

Hence $s_d=0.041$.

Substituting for u gives

$$u=\frac{\bar{x}-x_1}{s_d}=\frac{1.09-1.02}{0.041}=1.71.$$

He now uses Table III which gives the 5 percent and 1 percent levels of u for various values of n and d. Here n is the size of the sample which contains the suspected value, while d is the number of degrees of freedom on which s_d is based. In the present case $n=3$ and $d=16$.

TABLE II. Data available previously.

Set	1	2	3
1	1.11	1.07	1.10
2	1.17	1.15	1.19
3	1.20	1.23	1.16
4	1.11	1.15	1.25
5	1.06	1.10	1.00
6	1.03	1.10	1.04
7	1.07	1.01	1.06
8	1.06	1.01	1.08

The observed value of u, 1.71, is less than the value of u at the 5 percent level, 1.90. Hence he concludes that the suspected value 1.02 is not significantly outlying. In other words the deviation of 1.02 from the mean of the set of three measurements is easily explainable in terms of the precision of the measurement process. Hence, 1.02 is accepted into the fold of good measurements.

When past data are available, the u ratio may be computed and Table III used just as the r ratio and Table I were used for the case where no past data were available. The procedure outlined above for the two cases (a) and (b) may be followed just as before (using u and Table III instead of r and Table I).

CAUTIONS AND COMMENTS

(a) Obviously, if the experimenter *knows* by direct observation that a mistake has occurred he should reject the observation. The tests of this

TABLE III. Upper percent points of the studentized extreme deviate.
$u = (x_n - \bar{x})/s_d$ or $(\bar{x} - x_1)/s_d$

n \	α=5 percent							α=1 percent						
d	3	4	5	6	7	8	9	3	4	5	6	7	8	9
10	2.02	2.29	2.49	2.63	2.75	2.85	2.93	2.76	3.05	3.25	3.39	3.50	3.59	3.67
11	1.99	2.26	2.44	2.58	2.70	2.79	2.87	2.71	3.00	3.19	3.33	3.44	3.53	3.61
12	1.97	2.22	2.40	2.54	2.65	2.75	2.83	2.67	2.95	3.14	3.28	3.39	3.48	3.55
13	1.95	2.20	2.38	2.51	2.62	2.71	2.79	2.63	2.91	3.10	3.24	3.34	3.43	3.51
14	1.93	2.18	2.35	2.48	2.59	2.68	2.76	2.60	2.87	3.06	3.20	3.30	3.39	3.47
15	1.92	2.16	2.33	2.46	2.56	2.65	2.73	2.57	2.84	3.02	3.16	3.27	3.35	3.43
16	1.90	2.14	2.31	2.44	2.54	2.63	2.70	2.55	2.81	3.00	3.13	3.24	3.32	3.39
17	1.89	2.13	2.30	2.42	2.52	2.61	2.68	2.52	2.79	2.97	3.10	3.21	3.29	3.36
18	1.88	2.12	2.28	2.41	2.51	2.59	2.66	2.50	2.77	2.95	3.08	3.18	3.27	3.34
19	1.87	2.11	2.27	2.39	2.49	2.58	2.65	2.49	2.75	2.92	3.06	3.16	3.24	3.31
20	1.87	2.10	2.26	2.38	2.48	2.56	2.63	2.47	2.73	2.91	3.04	3.14	3.22	3.29
24	1.84	2.07	2.23	2.35	2.44	2.52	2.59	2.43	2.68	2.85	2.97	3.07	3.15	3.22
30	1.82	2.04	2.20	2.31	2.40	2.48	2.55	2.38	2.62	2.79	2.91	3.01	3.08	3.15
40	1.80	2.02	2.17	2.28	2.37	2.44	2.51	2.34	2.57	2.73	2.85	2.94	3.02	3.08
60	1.78	1.99	2.14	2.25	2.33	2.41	2.47	2.30	2.52	2.68	2.79	2.88	2.95	3.01
120	1.76	1.97	2.11	2.21	2.30	2.37	2.43	2.25	2.48	2.62	2.73	2.82	2.89	2.95
∞	1.74	1.94	2.08	2.18	2.27	2.33	2.39	2.22	2.43	2.57	2.68	2.76	2.83	2.88

* From K. R. Nair, Biometrika 35, 143 (1948).

paper are used only if he does not know that a mistake has occurred.

(b) If the experimenter uses this technique for a certain routine type of measurement, he should apply it, implicitly or explicitly, every time he makes that type of measurement. After several explicit applications of this technique, he will probably be able to perform the r (or u) test in all but the most doubtful cases without actually explicitly doing the arithmetic, since he will have the critical value of r (or u) in mind. He should not, however, reject outliers by the r test in some cases and accept others just as badly deviating, simply because he did not apply the test in these latter cases.

(c) Both the r and u tests are based on the assumption that repeated measurements of the same unknown follow the *normal frequency distribution*. If, in actual practice, the distribution of repeated measurements is markedly different from the normal curve, then the use of these tests will lead to different risks than originally intended.

(d) The use of the 0.01 and 0.05 points is arbitrary. The individual experimenter should use whatever *levels of significance* are most appropriate. It is accepted practice to choose the level of significance at the time the experiment is being planned and before any data are collected.

(e) Suppose the type of measurement is such that the suspected value will practically always be the smallest in the set, or practically always the largest. Then as stated above, 1 percent of the time a perfectly good observation will be rejected in case no additional observations are possible. Suppose, however, the type of measurement is such that the suspected value may be either the largest or the smallest. In this case about 2 percent of the time a perfectly good observation will be rejected. The appropriate tabular point should be selected with this in mind.

(f) Other tests[2] for rejection of suspected values are available. However, the r and u tests have been selected because of their ease of application.

BIBLIOGRAPHY

1. F. E. Grubbs, Ann. Math. Stat. 21, No. 1 (1950).
2. K. R. Nair, Biometrika 35, 118–144 (1948).
3. K. R. Nair, Biometrika 39, 189–191 (1952). (Contains additional *upper* probability points supplementing those of Table II.)
4. E. S. Pearson and C. C. Sekar, Biometrika 28 (1936).
5. P. R. Rider, "Criteria for rejection of observations," Washington University Studies, New Series, Science and Technology—No. 8, October, 1933.

[2] Dixon, Ann. Math. Stat. 21, No. 4 (1950).

6. Miscellaneous Topics

Foreword

One of the most delightful papers to read in this volume is D. B. De Lury's Computation with Approximate Numbers. This paper (6.7) explains once and for all the difference between computation operations with arithmetic numbers and computation operations with numbers resulting from measurements. No one engaged in the field of measurement can afford not to read it.

Edwin L. Crow's Optimum Allocation of Calibration Errors (6.4) considers the way errors accumulate in a hierarchy of calibrations, and proposes optimum allocation of errors within such a system, from the point of view of minimizing the total cost of achieving a given accuracy. Optimum error ratios were computed for several examples under extremely simplified assumptions. Much work needs to be done in this direction before the results can be fruitfully used.

R. B. Murphy is the author of ASTM E177-61T, Use of the Terms Precision and Accuracy as Applied to Measurement of a Property of a Material, and is currently working on a revised version to be issued as a standard. His paper (6.1), giving some background philosophy on the meaning of the terms precision and accuracy, was presented in the ASTM Symposium on Quality of Observations in 1961.

Two other papers, one by W. J. Youden and one by Milton Terry, are reprints from the same Symposium. Youden's paper (6.2) emphasizes, as he always does, the use of experimental design to throw light on the sources of errors. Terry (6.3) presents an example of the use of control charts on residuals. For further reading on the important subject of residual analysis, one may begin with Chapter 3, The Examination of Residuals, in Applied Regression Analysis by Draper and Smith (Selected References B8).

The relationship between confidence intervals and tests of significance, and the interpretation of confidence intervals and tolerance intervals, have always been sources of difficulty to some. Two papers, one by Mary G. Natrella, (6.6) and one by Frank Proschan, (6.5) are included here for the clarification of these concepts.

Quality of Observations*

YOUR test program is now complete and your file bulges with numbers. Two questions arise: How big are the numbers? How good are they? The following four papers address themselves to the second question. Here you will find definitions of those much-debated terms, "precision" and "accuracy," together with methods for determining them. Here you will also find suggestions on how to plan your experiments so as to improve the quality of your observations.

On the Meaning of Precision and Accuracy

By R. B. MURPHY

FOR SOME YEARS, the terms precision and accuracy have been used in connection with problems of measurement. About ten years ago ASTM Committee E-11 on Quality Control of Materials set itself the task of setting down some definitions for these two ideas. Their work on this subject is not completely finished even now. The words "accuracy" and "precision" have appeared in many places in ASTM standards and practices over the years. Other committees besides E-11 have attempted to set down standard definitions.

Debates and arguments about these terms seem to go on and on, so that the job of setting down definitions is a tough one. It is always a problem in defining ideas to balance rigor and exactness against practicality and simplicity; and in the present case matters have been made worse by a rather prolonged disagreement over which of two particular meanings the word "accuracy" should come to have.

The Measurement Process

Before we discuss the development of the E-11 definitions, I should like to adopt some terms for purposes of discussion. First and foremost, I should like to draw a distinction between a measurement or test method and a measurement process. A test method consists of a prescription or written procedure by which one can go about the business of making measurements on the properties of some physical material. This prescription may be very specific indeed, but essentially it is a much more inanimate object than a measurement process. A measurement process includes: (*a*) measurement method, (*b*) system of causes, (*c*) repetition, and (*d*) capability of control. A measurement process we could call a realization of a method in terms of particular men, particular equipment, and particular material to be tested. Of course, there is the question of whether a process is loyal to the method that it attempts to follow, or whether two different processes should be considered realizations of the same method.

It is handy here to import some of the language of statistical quality control to further characterize a measurement process. A measurement process may be regarded as a product of a system of causes, some of which may or may not have been specified in the test method. The important thing at this point is to recognize the broad scope of meaning embraced by the notion of a system of causes. A system of causes encompasses: (*a*) the material, (*b*) operator, (*c*) instrument, (*d*) laboratory, and (*e*) day.

Following through with this line of thought borrowed from quality control, we shall add a requirement that an

* The following four papers and discussion were presented at the Thirty-fifth Session of the Sixty-third Annual Meeting of the Society, held in Atlantic City, N. J., July 1, 1960. The symposium was jointly sponsored by the Administrative Committee on Research and Committee E-11 on Quality Control of Materials. A. T. McPherson, associate director, National Bureau of Standards, was symposium chairman.

NOTE—DISCUSSION OF THIS PAPER IS INVITED, either for publication or for the attention of the author or authors. Address all communications to ASTM Headquarters, 1916 Race St., Philadelphia 3, Pa.

R. B. MURPHY is a native of Massachusetts who has spent most of his life in the New York metropolitan area. He holds graduate and undergraduate degrees in mathematics from Princeton University with time out for service in the U. S. Marine Corps in World War II. After teaching mathematics and statistics at Carnegie Institute of Technology, he took up his present work at Bell Telephone Laboratories, Inc., New York, N.Y. on statistical problems arising in quality assurance.

Materials Research & Standards

April 1961

effort to follow a test method ought not to be known as a measurement process unless it is capable of statistical control. Capability of control means that either the measurements are the product of an identifiable statistical universe or an orderly array of such universes or, if not, the physical causes preventing such identification may themselves be identified and, if desired, isolated and suppressed. Incapability of control implies that the results of measurement are not to be trusted as indications of the physical property at hand—in short, we are not in any verifiable sense measuring anything. Of course, it is profoundly difficult to say how capability of control shall be ascertained.

There is, however, a relatively simple procedure or body of related procedures for substantiating—or even defining—a state of statistical control. If, in fact, we have statistical control—and not merely the capability of it—and if for some reason such control, however we gage it, appears to be lost, we would be ready, willing, and able to take some special action beyond that normally entailed in the test method alone. Such action would have the aim, of course, of restoring our confidence in the capability of the measurement process to be statistically controlled and, indeed, to restore such control, if possible.

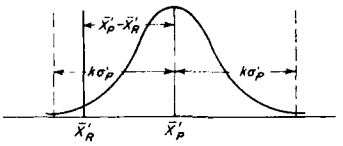
Why, one may ask, is there any need to impose the requirement of capability of statistical control? It is very simple. Without this limitation on the notion of measurement process, one is unable to go on to give meaning to those statistical measures which are basic to any discussion of precision and accuracy.

In any particular case, failure to have statistical control casts doubt on the sufficiency of our knowledge of the system of causes. It is then a question of determining which causes responsible for lack of statistical control should be acknowledged and included in our concept of the measurement process at hand and which should be eliminated so far as possible in their effect on the measurement process. Such elimination may entail a new prescription for the test method itself.

Reference Level or Target Value

One element of the system of causes which may be changed deliberately, although perhaps with unpredictable consequences, is what we may call the reference level of the quality of the material tested. A change of material would ordinarily imply a change in the reference level. This single cause in the system of causes has a unique position of importance in any measurement process. Some people prefer the term "true value," although others excoriate it as philosophically unsound.

We could also call the reference level a "target value." In a way this is a bad term because it implies that it is something we *want* to find through the measurement process rather than something we ought to find because, like Mt. Everest, it is there. Unfortunately our desires can influence our notion of what is true, and we can even unconsciously bring the latter into agreement with the former; my use of the term "target value" is not meant to imply that I think it legitimate to equate what we would like to see with what is there.



Precision is indicated by a multiple of σ'_P. $\bar{X}'_P - X'_R$ is called bias.

Fig. 1.—Precision and bias.

On the other hand, "target value" is a suggestive term (hopefully, not overly so) for purposes of present discussion. It is, in fact, interesting to compare the measurement situation with that of a marksman aiming at a target. We would call him a precise marksman if, in firing a sequence of rounds, he were able to place all his shots in a rather small circle on the target. Any other rifleman unable to group his shots in such a small circle would naturally be regarded as less precise. Most people would accept this characterization whether either rifleman hits the bull's-eye or not.

Surely all would agree that if our man hits or nearly hits the bull's-eye on all occasions, he should be called an accurate marksman. Unhappily, he may be a very precise marksman, but if his rifle is out of adjustment, perhaps the small circle of shots is centered at a point some distance from the bull's eye. In that case we might regard him as an inaccurate marksman. Perhaps we should say that he is a potentially accurate marksman firing with a faulty rifle, but speaking categorically, we should have to say that the results were inaccurate.

Components of Precision and Accuracy

One school of thought on the subject of accuracy insists that if a marksman hits the bull's-eye "on the average," then he is accurate even though the man may have a wavering aim so that his shots scatter. The point is that accuracy in this sense is determined solely by the behavior of the long-run average of the shots. The position of the average shot is assumed, of course, to be the centroid of the bullet holes in the target: few shots might actually hit or nearly hit the bull's-eye.

The second school of thought on accuracy would insist that if the man is unlikely to be very close to the bull's-eye he should be termed an inaccurate shot. That is, the second school holds to the belief that accuracy should imply that any given shot is very likely to be in the bull's-eye or very near to it. Both schools of thought have meaningful and verifiable versions of the comparatives "more accurate" and "less accurate," although if one follows the second school of thought, such a comparison is not always possible.

We may regard the rifle-range rules, the specifications of the rifle, ammunition and target, and manual for marksmen as analogous to a test method; the marksman and his rifle firing away at a specific target, on a specific range, perhaps on a specific day, correspond to a measurement process. Likewise, it is easy to translate the difference in viewpoints with regard to accuracy just noted from the field of marksmanship to the field of measurement and testing.

Before going further, we had best put down some elementary notions that we intend to use with respect to the problem of precision and accuracy in measurement The first of these is the long-run average of the measurement process, designated by $\bar{X}'_P$ (Fig. 1). It is assumed in this case that our measurement process produces a series of numbers and that therefore the quantity denoted by $\bar{X}'_P$ is a single real number. The reference level will be denoted by X'_R. The difference between these two quantities is almost universally referred to as "bias." Some have used the term "systematic error" synonymously, but others prefer to regard systematic error as the cause of bias. Another notion of primary importance is the standard deviation of the measurements produced by the measurement process. For this we have the symbol, σ'_P, and we regard this as a long-run characteristic of the process just as we do $\bar{X}'_P$. In words, the definition of the standard deviation is the square root of the mean squared deviation of the measurements from $\bar{X}'_P$.

Definition of Accuracy

Now let us return to our debate about the definition of accuracy. It is impossible to say that one of these viewpoints is wrong and the other is right from a sheerly logical point of view. I can put forth an argument relative to the conservation of linguistic resources. It seems to me that the terms "bias" and "systematic error" are adequate to cover the situation with

which they are concerned. If, nevertheless, we add the term "accuracy" to apply again in this restricted sense, we are left wordless—at the moment at least—when it comes to the idea of over-all error. From the point of view of the need for a term it is hard to defend the view that accuracy should concern itself solely with bias.

It is also important to determine whether one or the other of these definitions of accuracy has practical advantages over the other. I feel that there are certain circumstances in which one may be preferred and certain circumstances in which the other may be preferred. I doubt that one could show that there are substantially more situations in which one of these is appreciably more suitable than the other.

We are then left with the problem: If we are to have a single recognized definition of accuracy, on what basis other than that of need will we choose between these two, assuming that these are the only two possibilities we wish to consider? It would seem that the only basis for decision is a consideration of how the term accuracy is now used. It must be conceded that the school that believes that accuracy should connote the agreement between a long-run average of measurement process and the reference level is one of long standing among some experimenters. It can be argued, too, that it is easy to use accuracy in this way, because it is then possible to measure accuracy in terms of bias or systematic error. On the other side, a paper by Churchill Eisenhart of the National Bureau of Standards[1] has had considerable influence. The Bell Telephone Laboratories have used accuracy in his sense for some years.

We can also look at what practices are being followed with respect to the use of the word "accuracy" in different ASTM standards. There are a negligible number of cases in which accuracy is explicitly described in ASTM standards as a property of the long-run average. Usually there is no clear statement of which concept of accuracy is intended. In most of the standards in which accuracy is mentioned or discussed, precision is not mentioned or discussed, and vice versa. While the meaning and usefulness of the exact quantities given may be open to question in some cases, the obvious intent of these statements with regard to accuracy is that of an all-inclusive notion of error of measurement. Incidentally, in some instances the term precision has been used with regard to over-all error of measurement. At least one ASTM paper has explicitly taken this same view of precision. Seldom is bias or systematic error singled out in this body of literature. Thus there is overwhelming evidence that we need a term at least for the concept of over-all error.

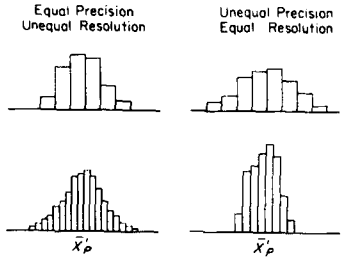


Fig. 2.—Resolution and precision (illustrated by frequency histograms).

On this basis I think there is considerable justification for the choice of Committee E-11 that accuracy should connote the idea of the error of individual measurements when that error is compounded of bias or systematic error and random or nonsystematic error.

Statistical Resolution

At this point I should like to inject one more note of confusion. It seems to me that one of the features of test methods which occasionally obtrudes itself in the arguments about definitions of precision and accuracy is the question of sensitivity and, as I shall call it, statistical resolution of measurement. Sensitivity sounds as though it ought to contribute to precision as we have described it. We could describe sensitivity as being measured by the minimum difference between the measurements of two different materials which we can possibly distinguish by the test method under consideration—the smaller the difference, the higher the sensitivity. Logically, if not conventionally, we might say that sensitivity should be the reciprocal of this quantity, but we shall follow the more conventional idea that sensitivity is directly measured by this minimum measurable difference.

At any rate, it is obvious that if our sensitivity is not very good, our precision is also not very good. However, the two are different, and we might define another quantity, to be called statistical resolution, which may be expressed as the ratio of sensitivity to standard deviation. If we can detect arbitrarily small differences in the property measured, the resolution is numerically small, because the sensitivity is small while the standard deviation is presumably a function of other factors as well which do not permit its magnitude to be arbitrarily reduced. Figure 2 exhibits some interesting distinctions that can be drawn. The left-hand pair of histograms in Fig. 2 have about equal spread, but the upper one is more coarsely grained, so to speak. Thus the sensitivity of a process producing the lower histogram would be greater than that of a process producing the upper one. Since the standard deviations are about equal, it follows that the resolution associated with the lower histogram is greater than that associated with the upper. On the other hand, if we spread the upper histogram out and squeeze the lower one together, as it were, without much change in the column widths, we should get something like the right-hand pair of histograms. The ratio of standard deviations would have been changed but not the ratio of sensitivities. If we spread and squeeze just the right amount, we can obtain equal resolutions although the sensitivities and standard deviations differ. This serves simply to emphasize that sensitivity is an absolute property and resolution a relative one in terms of the units of measurement. It may be useful to consider this kind of statistical resolution in measurement problems more than it has been thus far.

It may be perfectly possible that one process has higher resolution (numerically smaller) than another and yet is less precise. The number 2 represents a "worst possible" resolution, so to speak: it is that of a process in which we are able to observe either one or another of two values with equal probability. In general, we would expect the resolution of a process to be numerically smaller than 2. For practical purposes perhaps we should prefer resolutions on the order of $\frac{1}{2}$ or less.

Measures of Precision and Accuracy

Another purpose of the E-11 practice is to give a common set of terms for describing the measures or indexes of precision or accuracy stated in particular standards. This is not an easy job either. First of all, different fields have particularly favorite ways of expressing precision. Most of these measures are multiples of the standard deviation; it is not always clear which multiple is meant. It is possible, of course, that a single simple multiple might not do.

Some consider it unfortunate that precision should be stated as a multiple of standard deviation, since precision should increase as standard deviation decreases. Indeed, it would be more exact to say that standard deviation is a measure of imprecision. However, sensitivity, as we have previously

[1] C. Eisenhart, "The Reliability of Measured Values—Part I. Fundamental Concepts," *Photogrammetric Engineering*, June, 1952, pp. 542–554.

indicated, suffers from this logical inversion without hurt. Perhaps we can best avoid this by saying that standard deviation is an index of precision. The habit of saying "The precision is . . ." is deeply rooted, and there would be understandable impatience with the notion that standard deviation should be numerically inverted before being quoted in a statement of precision.

Some obvious choices of multiples of standard deviation for indexes of precision are given in Table I. The standard deviation itself, of course, may be used as an index. Sometimes the precision is stated as ±2 standard deviations with the implication that approximately 95 per cent of all the measurements of the measurement process will fall within two standard deviations of the long-run average for that process, whether that long-run average agrees with the reference level or not. In some cases people have used the multiple 1.96 rather than 2 in the hope that they will have obtained limits which more truly represent actual bounds within which 95 per cent of the universe will lie. Usually such refinements are specious on two grounds: first, because the accuracy with which the standard deviation will be known is not consistent with distinguishing between multipliers of 2.00 and 1.96; second, too great a reliance on the figure of 95 per cent is unjustifiable, anyway, since some measurement processes will yield a universe of observations of which perhaps only 90 per cent may lie within the 2-standard-deviation limits. It is reasonable to suppose in most cases, however, that such limits will include 90 to 95 per cent of the statistical universe of observations. Because of the uncertainty associated with this multiple, it might usually be better avoided in favor of other alternatives.

Precision is often stated as ±3 times the standard deviation, with the idea that for all practical purposes a measurement process, assumed to be under control, should be expected to yield measurements only within a 3-standard-deviation band about the long-run process average.

In some fields a preference has been shown for expressing precision not so much as a difference between an observation and the long-run average value of the measurement process but rather as a difference between any two observations from the same process. This has led to limits analogous to those previously mentioned and calculated from them by multiplying by $\sqrt{2}$. There is again a problem of giving such things names.

TABLE I.—INDEXES OF PRECISION.

Term	Reference Abbreviation	Notation
One-Sigma Limits	1S	$\pm\sigma'_P$
Two-Sigma Limits	2S	$\pm 2\sigma'_P$
Three-Sigma Limits	3S	$\pm 3\sigma'_P$
Difference Two-Sigma Limits	D2S	$\pm 2\sqrt{2}\sigma'_P$
Difference Three-Sigma Limits	D3S	$\pm 3\sqrt{2}\sigma'_P$

There are other distinctions to be made, however, which should be as clear as possible in any statement of accuracy. Frequently precision is stated as a percentage, such as the coefficient of variation. Any of the above indexes of precision can be converted to a percentage, but it is not altogether clear that there is only one figure of which these may be stated as percentages. Obviously the long-run average of the process is an outstanding candidate to use as a means of expressing percentage figures. However, this may not be convenient in all cases. In some areas it is not unusual to use a single fixed quantity of which precision is stated as a percentage.

Furthermore, the precision of a process may alter with the reference level regardless of the way in which we indicate the precision, whether as a standard deviation or a standard deviation expressed as a percentage of some other number. If that is so, the use of a single number on a standard then raises a question. Does this mean that the precision is constant over the range of reference levels in which we could possibly be interested or does this single figure of precision mean something else? Certainly it is not uncommon to consider this to be a maximum figure of precision over all possible levels of interest. If so, it would be well to append the word "max" after the stated precision of the process.

Again it is often desirable to qualify the statements of precision by some reference to the system of causes for which the statement of precision is valid. For instance, is this the kind of precision we should expect if we have one highly trained scientist operating one carefully adjusted instrument in a laboratory? Is it what we should expect over a short period of time or over a long period of time? Is it what we should expect of industry-wide comparisons of the same material? And so on. Such qualifying terms as "single operator," "interlaboratory," "single-day" are helpful to the interpretation of statements of precision. Perhaps even more important, thinking about these things is likely to be a big help in getting one to state the precision that he is really interested in in the first place. Sometimes we cannot succeed in being altogether explicit, but efforts to do so in this regard may very well help in the attainment of valid statements of precision.

TABLE II.—INDEXES OF ACCURACY

Term	Reference Abbreviation	Notation
Precision and Bias	...	$k\sigma'_P,\ \bar{X}'_P - \bar{X}'_K$
Limits of Error	LE	$\bar{X}'_P - \bar{X}'_E \pm 3\sigma'_P$
Root Mean Square Error	...	$((\bar{X}'_P - \bar{X}'_E)^2 + \sigma'^2_P)^{1/2}$

What has been said of precision can be said also of accuracy with regard to the terms and clarity of reference. The particular measures used are somewhat more difficult to deal with. This is because we have used the definition of accuracy which involves the combination of random and systematic error. Perhaps the most satisfactory way of expressing accuracy is to express precision in some way and then also to state the bias in a comparable manner. Both these figures could be represented as quantities which may vary as the system of causes is altered in some respects. This and other possible means are set down in Table II. The root mean square of error has nothing in particular to recommend it except statistical history. It cannot be used in any simple straightforward way, nor is it much help in efforts to visualize the situation with regard to experimental error. It has been dropped from the practice.

Thus, we hope this practice may provide a way of interpreting consistently and exactly such statements as "the precision of the method is ±2 per cent (relative per cent S.D.) max." Reference to this practice would, we hope, facilitate such consistent interpretation.

Verification of Precision and Accuracy

There is one very obvious problem, among others, which is not discussed at all in the recommended practice to be issued by ASTM Committee E-11. That is the problem of verification of the precision or accuracy of a measurement process. Anyone will acknowledge that assessing the precision or accuracy is a prerequisite to stating it. It is not so easy to see just how one goes about doing this. Other speakers at this symposium will discuss this subject. However, it is pointed out in the Recommended Practice that any such process of assessment is in itself a measurement process distinct from one that exists for the purpose of testing materials and evaluating them on a routine basis.

How to Evaluate Accuracy

BY W. J. YOUDEN

This paper presents a logical breakdown of the error in a measurement into (*a*) the systematic error inherent in the procedure, (*b*) the local systematic error of the laboratory using the procedure and, (*c*) the random error (precision). This breakdown should facilitate efforts to attain better accuracy. Several methods are given for identifying sources of error in measurements.

THE term accuracy conveys to most the idea of a value that is very close to the truth. The "truth" has to be defined rather carefully. Absolutely pure sodium chloride undoubtedly has a composition which conceptually, at least, corresponds to a certain weight per cent content of chlorine and a residual weight per cent of sodium. The presently accepted *atomic weights for chlorine and sodium* can be used to calculate the weight composition. This calculated result, admittedly, is not the absolute truth, but it has to serve in that role. A chemist, trying out an analytical procedure, will take this calculated composition as the truth.

Systematic Errors

Good agreement among repeat measurements in no way implies that the average of the measurements is close to the "truth" when the truth is some conceptual value of the property under measurement. Experience shows that averages of increasing numbers of repeat measurements, made under uniformly maintained conditions, do converge upon a particular value that reflects the true value but also depends in part upon the procedure, equipment, and environment used to make the measurement.

In the ideal situation the limiting mean that the averages of repeat measurements converge to would be the same as the true value. The difference between the conceptual true value and the average of the measurements is an estimate of the systematic error associated with the particular procedure and the circumstances providing the measurements. If there is evidence of a systematic error when the procedure is used in several laboratories, then this systematic error may be taken as a property—undesirable—of the particular procedure.

Some care is necessary at this point. Again, experience shows that if a measurement procedure is used at different times and places, that is, in different laboratories, the measurements converge to different average values. These average values are often maintained for considerable periods for the different laboratories and reflect inevitable differences in reagents and in the calibration of instruments; also differences between localities in humidity, temperature, etc., and finally some possible differences in the interpretation of the instructions for making measurements. Every round robin results in a collection of laboratory averages that differ among themselves by more than can reasonably be accounted for by the within-laboratory precision. Some point of view needs to be adopted toward the collection of systematic errors that are available when a value acceptably close to the true value is available for comparison.

One convenient viewpoint, whenever enough laboratories are involved, is to designate the average of all the laboratory averages as a grand average, characteristic of the procedure. The difference between this grand average and the true value can be considered an estimate of the systematic error of the procedure. The scatter exhibited by the individual laboratory averages suggests that calibration errors, and all other departures from the norm, introduce positive or negative departures from the normal systematic error of the procedure. There are two important consequences of this point of view. First, the difference between a laboratory average and the true value is not regarded as a single item but rather as a composite of two items, namely, the systematic error of the method modified by a systematic error of the laboratory as measured from the grand average. The second consequence is that the systematic error of the laboratory, relative to the consensus of all laboratories, can be obtained even when the true value of the property is not known. Even when the true value is known, it does not seem fair to charge a test laboratory with the systematic error that is an inherent property of the procedure as shown by the consensus of all laboratories. A test laboratory should be held responsible only for departures from the performance that the procedure is capable of giving. The consensus of the laboratories seems a reasonable appraisal of the procedure.

There is another interesting consequence of the concept of the procedure average. Figure 1 shows the averages for a chemical analysis for each of nine laboratories marked on a scale of values. Also marked is the procedure average (grand average of all laboratories) and the assumed true value computed from the atomic weights. The procedure average is about 0.2 per cent above the theoretical composition, and this may

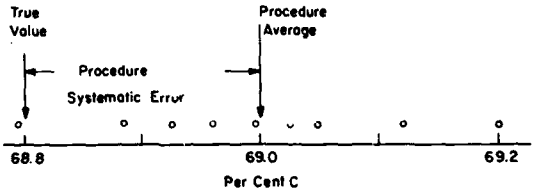


Fig. 1.—Averages for nine laboratories.

W. J. YOUDEN's academic degrees are in chemical engineering and chemistry. He began to use statistical procedures in 1925 when he was appointed chemist at the Boyce Thompson Institute for Plant Research, Inc. He held this post for 24 years except for the war period when he served as operations analyst with the Air Force. Since 1948 he has been a statistical consultant in the Applied Mathematics Division of the National Bureau of Standards, Washington, D. C.. Mr. Youden is the author of more than 100 papers, has written a book (*Statistical Methods for Chemists*), contributed statistical chapters to several other books, and for six years wrote a column "Statistical Design" for *Industrial and Engineering Chemistry*.

PERMISSION TO REPRINT OR REPUBLISH THIS PAPER ... other for publication or for ... author or authors. Address ... to ASTM Headquarters, 1916 Race St., Philadelphia 3, Pa.

be taken as an estimate of the systematic error of the procedure. The nine laboratories are scattered over a range of about 0.4 per cent. The lowest laboratory average is virtually coincident with the true value; the highest laboratory average is 0.4 per cent above the true value. Heretofore, the lowest laboratory (in this instance) would expect congratulations and the highest laboratory would be suspect. Quite the contrary interpretation can be made. There is no basis to consider either laboratory as doing better work than the other. Both laboratories have averages that depart by about equal amounts from the procedure average. Unless the low laboratory can describe some departure from the prescribed procedure to account for its result, no credit should be given for the accidental coincidence with the true value. Incidentally, if a departure from the procedure is admitted, this useful information should have been made available to the committee when the instructions were being prepared. Departures from the *procedure average* cannot be ignored. Indeed, unless and until a procedure has been adequately described, so that nearly all the laboratories show acceptable agreement for their averages, the question of agreement with a true value is hardly meaningful. If laboratories disagree, the procedure needs more careful specification. If the procedure average differs by an unacceptable amount from the true value, the procedure itself requires modification or rejection.

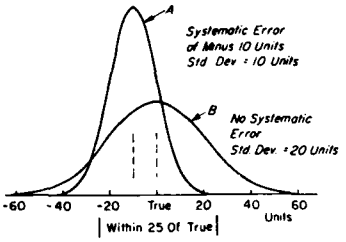


Fig. 2.—Which one is the more accurate?

It is worth noting that the usual evolution of a procedure does not suggest the viewpoint discussed above. Generally a particular laboratory works out a procedure and, because it gets highly satisfactory results, urges a trial by other laboratories. If this procedure happened to have been first tried by the laboratory that got the highest result in Fig. 1, perhaps nothing more would have been heard of the procedure. If the lowest laboratory in Fig. 1 was the first to try, then this laboratory becomes an enthusiastic sponsor of the procedure. One cannot escape the evidence that the laboratories are spread out and that any one of them might have been the originator of the procedure. There is a possibility that, for the sponsoring laboratory, a chance combination of instruments environments, etc. approximately canceled out the inherent systematic error of the procedure. Confusion will reig.. until the evidence is reviewed in the proper light. The procedure reported in Fig. 1 does have a systematic error as shown by the fact that eight of the nine laboratory averages have positive deviations from the true value.

Precision and Accuracy

There is much evidence that the systematic errors of laboratories, even when measured from the consensus, often tend to be as large or larger than the standard deviation computed for the random deviations associated with the precision. Even rather small systematic errors are fairly easy to demonstrate, because the random error of an average is inversely proportional to $\sqrt{n}$, where n is the number of measurements in the average. Consequently, a relatively few measurements stabilize reasonably well a laboratory average. In passing, it should be remarked that a much larger number of measurements are necessary to obtain a good estimate of the precision. Fortunately, the precision appears to be much the same for most laboratories using a procedure so that a pooled estimate of the precision is usually employed.

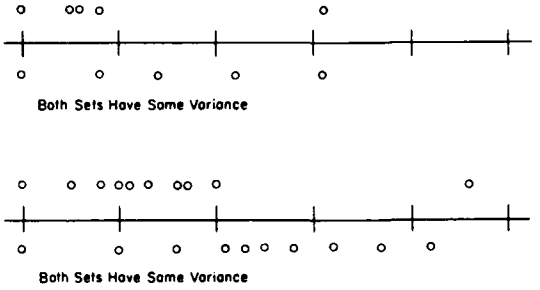


Fig. 3.—Variance does not tell the whole story.

One question is often raised. Is a procedure with a small systematic error to be preferred to one with practically no systematic error if the latter has much poorer precision? Suppose a precise procedure, *A*, with a standard deviation of 10 units has a systematic error of 10 units. Still, 93.3 per cent of individual results will be within 25 units of the true value. Another procedure, *B*, without systematic error, but with a standard deviation of 20 units, will have only 78.9 per cent of the individual results within 25 units of the true value. The error curves are shown in Fig. 2, and clearly more of the area of curve *A* lies within 25 units of the true value. So, on a single result, there is a better chance of an error less than 25 units using the procedure with the systematic error. Some writers have suggested that, from this point of view, *A* is the more accurate procedure.

This advantage of the more precise procedure does not always apply. Consider a manufacturer shipping many lots of his product. If the manufacturer is paid on the amount of active ingredient in his product, he will lose money in the long run using procedure *A*. His average will be 10 units lower than it would have been if procedure *B*, had been used. True, the results will fluctuate more with procedure *B*, but the losses and the gains will, in the long run, cancel out. This manufacturer no doubt would regard procedure B as the preferred procedure.

The Evaluation of Accuracy

There is no solution to the problem of devising a single number to represent the accuracy of a procedure. All through the preceding discussion accuracy has been associated with the test procedure rather than with the numerical measurement that results from using the procedure. The performance of the test procedure has to be established, and, barring evidence to the contrary, the measurements obtained by the procedure are considered to be subject to a particular systematic error and to have a particular precision. By various devices the systematic error may be allowed for. In routine work a reference specimen often permits the introduction of a correction that simultaneously adjusts for both the procedure systematic error and the local systematic error.

A more troublesome matter concerns the desire to attach to a reported result some statement of confidence limits for the result. The question is sometimes put in the form: What confidence limits apply to a result reported by a new laboratory not included in the original group participating in the study of the procedure? There is a pitfall here that will catch some who have uncritically accepted certain statistical techniques. One may glibly say that there is a certain within-laboratory

error and, in addition, a between-laboratories error that need only to be combined. The upper half of Fig. 3 shows two hypothetical sets of laboratory averages. Both sets have five laboratories, and both have the same between-laboratory variance. Nevertheless the two sets correspond to substantially different situations. In the set below the axis the laboratories fall into a reasonable pattern that might, conceivably, arise if the laboratory systematic errors were normally distributed. The set above the axis shows one extreme laboratory with the others compactly grouped and conveying a picture of a far more satisfactory *procedure*. More laboratories emphasize this contrast (Fig. 3, lower half).

In predicting what may happen if a new laboratory is included, statistical formulas lead to the same result for both sets. Few experienced laboratory workers will feel comfortable at this equivalence. More likely these workers would be inclined to get the extreme laboratory in the upper set to locate its trouble or else drop it from the group. Indeed, is it fair to judge the procedure with this laboratory included? The matter of confidence limits rests upon the presumption of a statistical distribution. Blind application of statistical formulas without thoughtful examination of the results may lead to absurd predictions.

The successful application of statistical methods rests upon a thorough understanding of the way the data were obtained. For example, a dozen repeat measurements made in close succession provide an estimate of the random variation to be encountered under such relatively unchanging conditions. If the dozen measurements are made one-by-one on randomly selected days over a period of weeks, the variation is usually larger. This additional component of variance can be merged with the within-day component. But if there is an awareness of a systematic error that applies to *all* the measurements, any well informed estimate of, say, the maximum size of this systematic error, must not be combined with the random component. Probability statements cannot be made about such combinations of random and systematic errors.

Detection of Systematic Errors

The differences so often found between the averages reported by several laboratories testify to the presence of systematic errors for at least some of the laboratories. Within one laboratory, other means are required to reveal any systematic error in the procedure. There are three major devices commonly used to test a measurement procedure:

1. Measurement of known materials.
2. Comparison with other measurement procedures.
3. Comparison with modifications of the given procedure.

More often than is realized a true value is known. All target shooting is a class of known values. The center of the bulls'-eye, or the assigned target coordinates in a bombing mission, is a known value. The objective is to hit the center of the target. The result of each aiming is a measurement usually reported directly as the "aiming error." Reflection shows that, given a collection of impact points, it will be more informative to locate, first, the centroid of the impact points. The displacement of this centroid from the assigned coordinates of the target corresponds to a systematic error, and the scatter of points about the centroid reveals the precision. Quite different steps will be needed to correct for the displaced centroid and to reduce the scatter about the centroid.

Often, in analytical chemistry, samples of known composition can be prepared. Spectrographic procedures are sometimes tested on materials analyzed by the more tedious and accurate "wet" methods of analysis. The "true" values thus established are often quite adequate for testing the spectrographic procedure. The standard materials prepared by the National Bureau of Standards are also used to provide materials with known "true" values.

Experimenters have long felt more at ease when two or more quite different procedures show agreement. Agreement does not prove the absence of a systematic error, but it does constitute evidence against the presence of a systematic error. Analytical chemistry offers many opportunities to try, on the same material, two or more analytical procedures that differ in the chemical reactions and reagents involved.

Another method, not used as often as it might be, makes use of a proportional relation when this exists. Consider a stock of material submitted to analysis. If several samples each weighing 2 g are analyzed, good agreement does not rule out the presence of a systematic error in all the results. But if samples of 0.5, 1.0, 1.5, 2.0, and 2.5 are tested, the weights of precipitate, or the volumes of reagent used should be strictly proportional to the sample weights. A straight line through the origin should fit the points if the observed results are plotted against sample weights. If there is a systematic error, constant over the range of sample weights, the points will be fitted by a line that intercepts the y-axis at a point corresponding to the systematic error. If the systematic error is proportional to the sample weights, the line will still go through the origin and the systematic error will not be revealed.

Test procedures for many materials lead to results which are not invariant under, for example, changes in specimen dimensions. Extremely careful specification of the test specimens is then necessary. The results are considered to be closely correlated with important properties of the material in bulk. Thus a cube of cement 2 in. on each edge may be submitted to a compression test. The results of such tests are used to determine whether the product meets certain specifications. Compression tests on cubes 3 in. on each edge could also be used, but, presumably, the relation of breaking load to cube dimensions is not a simple one.

Refined measurements of certain physical constants usually have systematic errors considerably in excess of the precision error attached to the average. Here an extremely carefully constructed set of equipment tends to give a series of readings showing superb agreement. Later, another worker, with an entirely different ensemble but based on the same principle, obtains an average unquestionably displaced from the results of preceding workers. Standard practice calls for the most painstaking elimination of sources of systematic errors often by introducing various corrections. Suppose, in the equipment, a tube of 1 mm in diameter is needed. An estimate will undoubtedly be made of the uncertainty introduced in the final result by the estimated uncertainty in the tube diameter. Surely the use of a second similar tube, or even one somewhat bigger or smaller, will provide an opportunity to estimate the effect of uncertainty in the tube diameter.

Experimenters immediately object that such dualization of each part of the apparatus would vastly increase the program. That is true. It is also true that a later investigator usually changes nearly everything. He gets a somewhat different result and there is no way to locate the reason. If the first man had tried two diameters of tube, and the second worker tried some other alternatives, then eventually there would accumulate the necessary information to pin down the source or sources of discrepancies.

Detection of Errors by Designed Experiments

Testing laboratories that run many tests of the same kind often overlook opportunities to check up on their

TABLE I.—SCHEDULE FOR PLACEMENT OF BARS.

Comparison Number	Bar Position East	Bar Position West	Difference, East-West	Comparison Number	Bar Position East	Bar Position West	Difference, East-West
1.....	A	B	d_1	6.....	A	C	d_6
2.....	B	C	d_2	7.....	C	E	d_7
3.....	C	D	d_3	8.....	E	B	d_8
4.....	D	E	d_4	9.....	B	D	d_9
5.....	E	A	d_5	10.....	D	A	d_{10}
	Total.....		Σd		Total.....		Σd

equipment without in any way interfering with their regular program of work. Two examples will be given, one a precision procedure and the other a more approximate measurement.

Meter bars are sometimes compared by placing them end-to-end in a long chamber. Every effort is made to maintain a uniform temperature the length of the chamber, otherwise spurious differences in the lengths of the bars may be introduced. Careful measurements are made to check on the uniformity of the temperature. The bars are intercompared in sets, every bar being matched with every other bar. A set of five bars makes possible ten pairings and consequently ten comparisons. Each comparison leads to a difference in lengths between the two bars in the chamber. Let one end of the chamber be designated the east end and the other end the west end. If the various pairs of meter bars are placed in the chamber without any plan, an opportunity for an easy test of the equipment will be lost.

One device long used to compensate for position effects is to reverse the positions of the objects and repeat the measurement. An alternative device achieves the same effect without actually reversing the positions for each pair. The objects may be scheduled for the positions so that, over the total of all the pairings used, each object will occupy each position the same number of times. The schedule shown in Table I has been used for this purpose. The letters, A, B, C, D, and E are used to identify the bars and the d's with subscripts denote the observed differences, the difference always being the length of the bar in the east end of the chamber minus the length of the bar in the west end.

Examination of the schedule (Table I) shows that the placement of the bars the chamber is such that, for the first five comparisons, all five bars have been in the east end and the same five bars also in the west end. When the five differences are summed this amounts to subtracting the total length of the five bars from the total length of the same five bars. The sum of these five differences should, therefore, be zero, within the limits of the measurement error. Suppose, however, that one end of the chamber is persistently slightly warmer than the other end. This will increase the length of the bars in the warm end and introduce a small bias in every observed difference. The sum of the five differences provides a very sensitive measure because the total length of all five bars is involved. The second set of five comparisons provides a check on the first result.

Two advantages accrue from such a planned assignment of bars. First, a temperature gradient may be detected or, it the sum of the differences is satisfactorily small, the evidence of position equality has been provided at no cost. Second, if there is a position effect the correction of the observed differences using the estimate of the systematic error, $\Sigma d/5$, is a simple matter.

Consider a piece of equipment with eight test positions. Perhaps duplicate specimens are usually run. In any event duplicate specimens of each test material will be needed for the 28 materials tested in the program in mind. With duplicate specimens and eight test heads, four materials can be compared in any run. Comparisons among materials, within a run, rely on the equivalence of the various test positions. The choice of the two positions assigned to the duplicate specimens can be used to throw light on the equivalence of the eight heads. Number the heads 1 to 8, the runs by Roman numerals, and the 28 materials by $a, b, \ldots, z, A, B$. The schedule for the assignment of duplicates of these 28 materials is shown in Table II. Each pair of duplicate specimens provides a difference. These differences should be entered in Table III by placing in each cell the difference obtained by subtracting one duplicate from the other in the order indicated. For example, material a tested in the first run gives the difference between positions one and five. Thus, in the first column the differences are obtained by subtracting from the result obtained on head 1 the appropriate results obtained on the other heads. The first-row entries list the same values with opposite sign.

The totals at the foot of the columns when divided by 8 rank the eight heads with reference to zero. As an arithmetical check, the sum of the eight column totals must be zero. A statistician's help will be useful in a complete analysis of these data. The experimental design presented here is intended to fit into the regular testing procedure with a minimum of interference. A simple direct way to compare the heads, in a special test, is to use eight specimens of the same material in a single run. About four such runs will be required to obtain as much information regarding head differences as is here obtained with 28 pairs of duplicates.

TABLE II.—SCHEDULE TO TEST EQUIVALENCE OF MACHINE HEADS.

Run Number	Head Number 1	2	3	4	5	6	7	8
I...	a	b	c	d	a	b	c	d
II...	e	g	g	e	h	f	f	h
III...	i	j	k	l	j	i	l	k
IV...	m	n	m	n	o	p	o	p
V...	q	r	s	t	s	t	q	r
VI...	u	u	v	v	w	w	x	x
VII...	y	A	B	z	z	B	A	y

TABLE III.—ARRANGEMENT OF DUPLICATE DIFFERENCES TO EVALUATE THE HEADS.

Head Number	Head Number 1	2	3	4	5	6	7	8
1.....	...	2-1	3-1	4-1	5-1	6-1	7-1	8-1
2.....	1-2	...	3-2	4-2	5-2	6-2	7-2	8-2
3.....	1-3	2-3	...	4-3	5-3	6-3	7-3	8-3
4.....	1-4	2-4	3-4	...	5-4	6-4	7-4	8-4
5.....	1-5	2-5	3-5	4-5	...	6-5	7-5	8-5
6.....	1-6	2-6	3-6	4-6	5-6	...	7-6	8-6
7.....	1-7	2-7	3-7	4-7	5-7	6-7	...	8-7
8.....	1-8	2-8	3-8	4-8	5-8	6-8	7-8	...
Total.....	Σ_1	Σ_2	Σ_3	Σ_4	Σ_5	Σ_6	Σ_7	Σ_8

Summary

The number of test procedures grows daily. The variety of equipment defies enumeration. Always the question of the sources of variation arises when test results show poor agreement. The written instructions for conducting tests contain phrases such as "shake vigorously," or "clean thoroughly." Operators will vary in the way they follow such instructions. Often no effort has been made to ascertain how vulnerable a test procedure is to moderate variations in the actual manual operations involved. Usually, if some major source of experimental variation can be located, steps may be taken to improve the situation. Fortunately, for every interesting test situation some equally interesting experimental design can be devised to throw light on the sources of experimental source. As these sources are identified and corrected the accuracy of test results will likewise improve.

On the Analysis of Planned Experiments

By MILTON E. TERRY

Over the past decade, scientists and engineers have increased the scope of their experimentation and the volume of test data to such an extent that additional analytic and reduction techniques have been required. With automatic data recorders of analog and digital types becoming almost commonplace, and with continuing enlargement of the body of scientific knowledge, it has become increasingly difficult for an experimenter to extract a satisfactory amount of information from his experiments.

The experimenter is finding himself more and more in the situation of the manufacturing or process engineer with far more data than he can ingest, digest, or understand.

It is not surprising, then, that several statisticians have returned to the pattern concepts of Shewhart and the other engineers interested in control of processes. Tukey and Anscombe,[1] and others have proposed several distinct and ingenious graphical techniques appropriate to various aspects of data analysis. Presented here are the technique and concepts I proposed and described.[2] This choice is personal and not dictated by scientific demand.

Over the past 30 years, two theoretical approaches to the statistical treatment of research and development problems have evolved. It is the purpose of this paper to show how both can be used together in the analysis of data.

W. A. Shewhart[3] and others have considered the problem of analyzing process data where the number of measurements is large. The approach proposed by Sir Ronald A. Fisher is to select a group of variables and a set of values of each variable, and then take measurements at selected combinations of these values. Then an estimate is made of the effect of changing each variable among its selected values, this effect being averaged over the selected values of each of the other variables. Randomization is used to average out the effects of the variables not under study.

The Shewhart method of analyzing data uses graphical methods wherein the data are first plotted in the pertinent recorded order in rational subgroups, and the applicable control limits found from an average "within-subgroup" estimate of dispersion. A subgroup central value and a dispersion estimate are plotted on charts together with their appropriate control limits. It is then standard practice to scrutinize all the charts for evidence of nonrandomness and lack of control. When the data finally pass all the tests of interest, estimation is justified. Of course, all datum points and statistics not satisfying a test criterion must be examined carefully by the research team for assignable causes. When the process yielding the data is not in control, estimation and prediction are hazardous.

Shewhart has pointed out that one may find sets of data which satisfy all simple statistical tests but display recurrent patterns which cast doubt on any hypothesis of randomness and independence. One of the most common

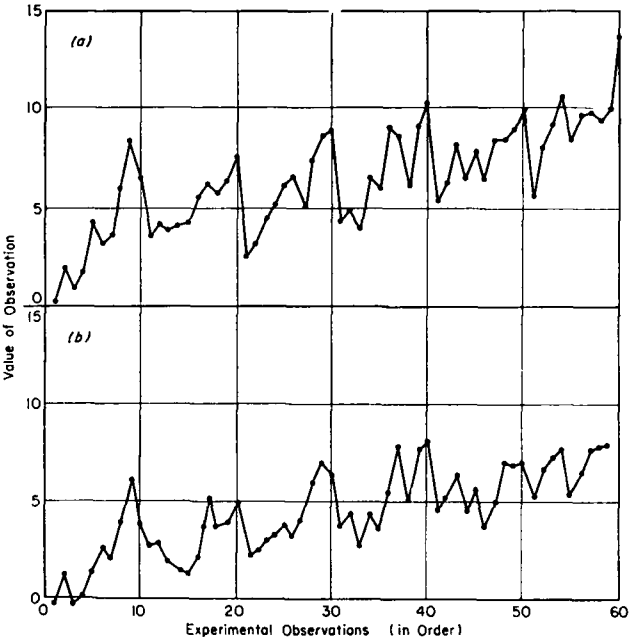


Fig. 1.—Experimental observations.

NOTE—DISCUSSION OF THIS PAPER IS INVITED, either for publication or for the attention of the author or authors. Address all communications to ASTM Headquarters, 1916 Race St., Philadelphia 3, Pa.

[1] F. J. Anscombe and J. W. Tukey, "The Criticism of Transformation," unpublished manuscript, 1954.

[2] M. E. Terry, "On the Analysis of Planned Experiments," *Transactions*, Am. Soc. Quality Control, pp. 553–556 (1955).

[3] W. A. Shewhart, private communication.

patterns he has found occurs in the field of multiple readings, forming trend lines of varying length and magnitude of slope, with sharp breaks between segments. When the variation of these lengths and slope magnitudes is small, certain inferences can be made. When the variation is large, it is not clear what inferences should be made or with what confidence.

The analysis of a statistically designed experiment using the classical form of the analysis of variance depends on three basic assumptions of (1) additivity of treatment effect, (2) independence, and (3) homoscedasticity. Under these assumptions it is possible to incorporate into almost all research projects a schedule of measurements on specified elements of the experiment involving the selected variables in such a way that the effects of each selected variable averaged over the combinations of selected values of the remaining variables can be measured. In addition, the reality of effect from a selected variable can be tested *statistically*. In fact, the testing of apparent reality of effect and estimation of residual variation have been the main functions of the analysis of variance, and until recently were considered a satisfactory ending to the reduction of experimental data. Hence, some engineering and industrial research personnel have cast aside the statistical design of experiments, since they could neither satisfy all of the assumptions nor accept the *classical form of the analysis of variance* as satisfactory at the end of most experiments where several or all of the following questions must be answered:

1. Are there any *assignable* causes of variation present other than those introduced into the experiment deliberately?
2. How important are the effects of each of the selected variables?
3. Was the experiment well conducted?
4. Were there any unusual outcomes worthy of study?
5. How large a fluctuation can be expected in the process for manufacturing a product of which the experimental units were originally presumed representative?
6. What specifications can be written?
7. Which of the selected variables have effects demonstrated by this experiment not to be zero?

The control chart technique gives answers to these questions, but not all have the same efficiency. The analysis of variance originally seemed to be designed to answer question 7 only, but with the aid of recent developments (components of variance, multiple comparisons, and the analysis of residuals) now offers reasonable answers to the remaining questions.

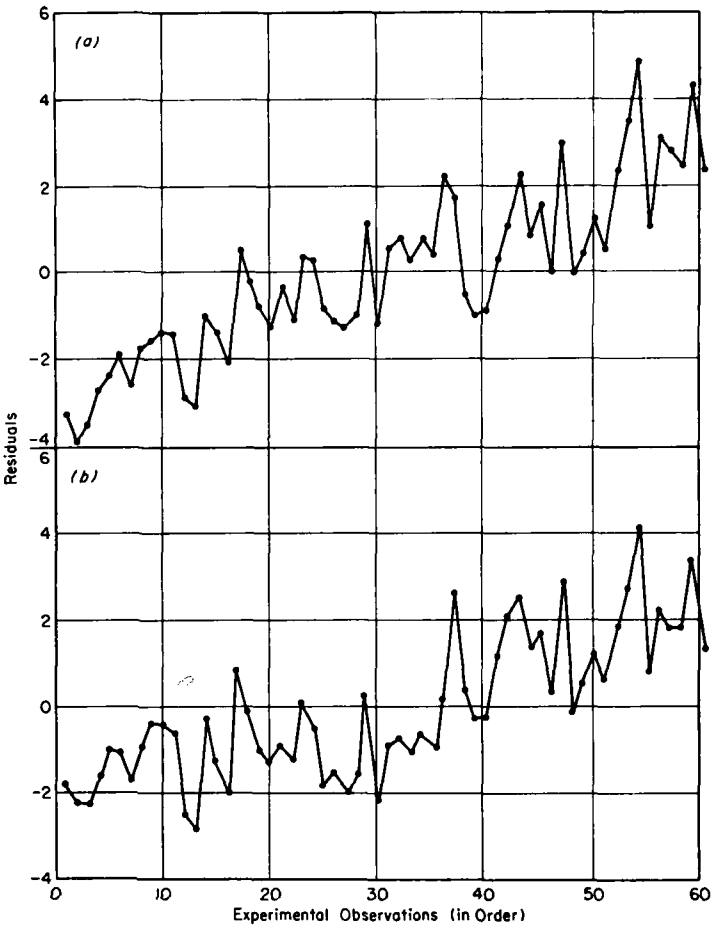


Fig. 2.—Residuals in order of manufacture.

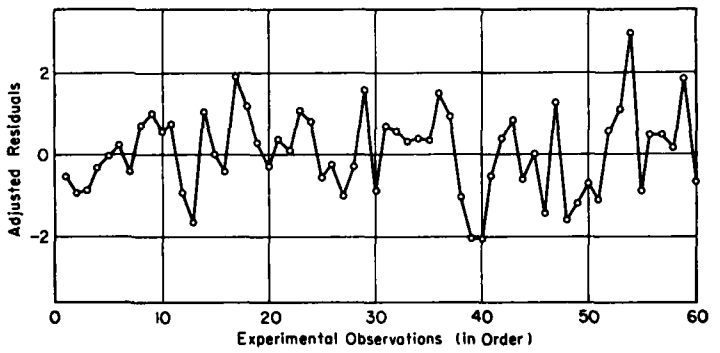


Fig. 3.—Adjusted residuals.

Under the assumptions of a statistically designed experiment we can always state a mathematical model. Consider the following hypothetical simple experiment. We wish to study the effect of reducing corrosion by evaporating a metal p mils in thickness on an electrical element. Ten elements at each of six thicknesses ($p_1, \ldots, p_6$)

are considered necessary. Only one element at a time can be coated, so the 60 units will be processed in a random order. They are to be subjected to a controlled corrosion attack and then measured. Let t_i be the true relative effect of thickness p_i in reducing corrosion ($\sum_{i=1}^{6} t_i = 0$). Let μ be the true average corrosion effect over the experimental range and y_{ij} the measurement of the jth element with the thickness coating p_i. Then our mathematical model is

$$y_{ij} = \mu + t_i + e_{ij}; \qquad \begin{matrix} i = 1, \ldots, 6; \\ j = 1, \ldots, 10 \end{matrix}$$

where e_{ij} is the residual effect and is assumed to be a random independent normal variate.

We can estimate μ by the over-all mean $\bar{X} = \sum_{ji} y_{ij}/60$; and t_i by $\bar{X}_i - \bar{X}$, where $\bar{X}_i = \sum_{j} y_{ij}/10$. Then we define $Y_{ij} = \mu + t_i$, $(i = 1, \ldots, 6)$ to be the predicted value, and $z_{ij} = y_{ij} - Y_{ij}$ to be the *residual* of the measurement ij. It follows that $\sigma^2 = \sigma^2_{z_{ij}} = S^2_{z_{ij}}/54$.

We stimulated this experiment by assigning constants to the μ and t_i and values to the ϵ_{ij} from a table of random order. In two simulations with respect to the ordered y'_{ij}, a linear trend and an abrupt shift in level were superposed respectively on the y'_{ij} to yield two sets of data y_{ij} of known behavior (see Figs. 1 (*a*) and (*b*)). Standard analyses were run. The estimates of relative mean effect were not very biased, but the estimates of the residual variation were so bad that no conclusions about equality of effects could be drawn. Then the z_{ij} were calculated for each simulation and plotted against order (see Figs. 2 (*a*) and (*b*)). When the data of Fig. 2 were corrected for the fitted trend line, the new estimates of the known parameters were excellent. The use of Fig. 3 gives an excellent estimate of the shift in level, and again correctly adjusted the estimates from Fig. 2

When the set of residuals, z_{ij}, constitute a time sequence, they can be plotted as such. In many engineering experiments, only one fabricating or measuring device is available, and hence one or more time sequences are imposed on the experiment. In general the statistical design will average out the time effect in the estimates $\hat{t}_i$ by randomizing the order of fabrication or measurement of the experimental units.

In a real sense, the set of residuals plotted against time, together with control limits, $\pm k_{\sigma_{\text{residual}}}$, are a control chart. Hence we are tempted to use the usual chart techniques. Since there may be constraints imposed by the model, the significance levels may be no longer identical with the tabular values. But when the control limits are used as action limits, satisfactory results should ensue.

Anscombe and Tukey have proposed plotting the set of residuals, z_{ij}, against its associated predicted value, Y_{ij}, when the experiment contains at least a double classification. Here "nonadditivity is shown by a curved regression. Nonconstancy of variance is shown by a wedge shape."

In general, plotting residuals both against their predicted values, and against serial order, s, enables the experimenter to examine that portion of his measurements which is not attributable to the suspect variables. He will have visual evidence as to the vexations from many sorts of nonadditivity of effect, nonconstancy of variance, linear trends, cycles, and wild shots which may be embedded in his experiment. Hence, the analyst-experimenter can take the necessary action to ensure that the final accepted readings in the proper units satisfy the assumptions on which valid predictions and estimates will be made. This form of analysis, used in conjunction with the analysis of variance, enables the user of a statistically designed experiment to focus the same type of scrutiny on his data that the control engineer can give to process data.

Optimum Allocation of Calibration Errors

EDWIN L. CROW

National Bureau of Standards, Boulder, Colorado

Answers are given to two questions, with emphasis on the second. (a) How do the errors accumulate from echelon to echelon in a hierarchy of calibrations? (b) If a certain accuracy is required at the final echelon of a hierarchy, what is the best way to achieve that accuracy, or, more specifically, what is the *optimum allocation of errors among* the echelons? The criterion for optimization is taken to be the minimization of the total cost of achieving a given accuracy.

Introduction

Since the art of measurement began there have been standards, more or less informal, by means of which further measuring sticks, weights, and capacity measures have been produced for use in construction and commerce.* With each reproduction of the measures variations were inevitably introduced, and these often consisted of intentional as well as accidental errors. The ancient Egyptians, Greeks, and Romans had respected standards of measure, but these fell out of use during the Dark Ages, and the later attempts to establish widely used standards were long doomed to failure.

In 1830 the United States Senate noted that variations in the standards in use at various customhouses were causing loss of revenue and directed the Secretary of the Treasury to make comparisons of these standards. The Treasury in fact took steps to supply uniform weights and measures to all customhouses, and the Secretary reported in 1832 that standards were being fabricated at the United States Arsenal in Washington "with all the exactness that the present advanced state of science and the arts will afford." Thus the Office of Weights and Measures came to be established in the late 1830's within the Treasury Department. In 1901, when its budget was still less than $10,000, the Office became a part of the new National Bureau of Standards. In 1903 the Bureau was transferred to its present position in the Department of Commerce.

Now the Bureau maintains hundreds of national standards and calibrates the standards of the states, military departments, manufacturers, utilities, universities, private testing companies, and others. The Bureau is unable to calibrate all secondary standards and instruments, and the above types of organizations in turn calibrate further standards. For example, counties and cities may have their balances, weights, and other measures certified by their state offices, and they in turn certify the balances within their jurisdictions.

In electrical energy the Bureau uses a standard watthour meter accurate to about 0.03 percent to calibrate the master standards of public utility commissions and power companies. The latter in turn make measurements to about 0.1 percent of customers' meters. As a result in part of variability in time, customers' meters operate within about one percent accuracy.[4]

In recent years the demanding requirements of missiles, spacecraft, and other vehicles have led to the establishment of extensive hierarchies of standards laboratories by the military departments. As indicated in Fig. 1, the National Bureau of Standards is at the apex of these hierarchies. The figure indicates just a few examples of the standards laboratories that enter in various levels, or echelons, of the hierarchy. For most basic standards the Bureau is itself just one of the many national laboratories deriving their units from the International Bureau of

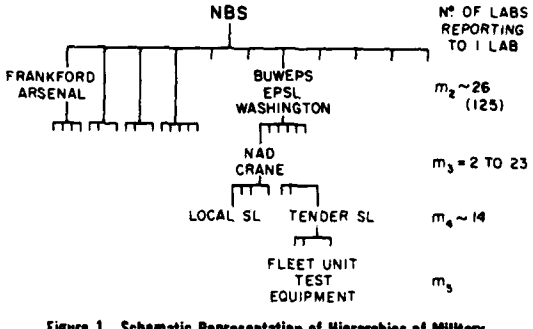


Figure 1—Schematic Representation of Hierarchies of Military Standards Laboratories Using National Bureau of Standards Calibration Services.

Dr. Crow is Consultant in Statistics, Environmental Science Services Administration, Boulder, Colorado.

Adapted from a paper delivered to the 18th Midwest Quality Control Conference on October 12, 1963, in Tulsa, Oklahoma.

*The introductory historical remarks are derived from the fascinating histories of standards written by John Perry[5] and Ralph W. Smith[6].

ASQC LCS Code 767:60;70:400

Weights and Measures. In each echelon of the hierarchy and with each transfer of information, some error is unavoidably introduced.

Two questions arise:

(a) How do the errors accumulate from echelon to echelon in a hierarchy of calibrations?

(b) If a certain accuracy (or, alternatively, precision; see Eisenhart[2] for a basic discussion including definitions) is required at the final echelon of a hierarchy, what is the best way to achieve that accuracy, or, more specifically, what is the optimum allocation of errors among the echelons?

The first question was discussed recently by Woods and Zehna[7], but the present article will summarize some complementary results, as well as mathematical and numerical answers to the second question. The basic material is drawn from a 1960 paper[1], in which more details may be found. It has often been stated that each echelon should be 10 times as accurate as the next one, but the answer to the second question will show that usually nowhere near that accuracy is required.

If we are to answer a question of the "best way" or "optimum" quantitatively, we must adopt some criterion for judgment. Here we have quite naturally adopted the criterion that, for a specified final accuracy (total error), the cost should be minimized. It is a pleasant fact that the answer would be the same if the criterion were that, for a specified cost, the final accuracy should be maximized.

How Do Errors Accumulate in a Hierarchy?

Before we answer the first question we must define "error" a little more precisely. Also, in order to answer this and the next question without unnecessary complication we shall make the simplifying assumption that the characteristic errors, however they are defined, and the characteristic costs of all calibrations in the same echelon of the hierarchy are the same.

Let us suppose that in a particular type of calibration there are n echelons, numbered from 1 to n starting with the top laboratory. A laboratory in the jth echelon of a hierarchy adds an error, say e_j, to the error passed on to it from laboratories higher in the hierarchy; e_j is an individual error, varying from day to day and calibration to calibration, perhaps positive, perhaps negative, and not known in an individual case. However, if there are n echelons, we can say that the total error of a measurement in the nth echelon (relative to the international standard), say e_{tot}, is given by the equation

$$e_{tot} = e_1 + e_2 + \ldots + e_n$$

Let us define E_j as the "maximum" numerical value of e_j, or, if we need to be more precise, the value that is exceeded by e_j numerically just 0.3 percent of the time in the long run. Likewise let E_{tot} be the maximum value of e_{tot}. Consider the extreme case in which the errors e_j and e_{j+1} in successive echelons are perfectly positively correlated, so that

$$E_{tot} = E_1 + E_2 + \ldots + E_n$$

If we assume that the error ratio E_{j+1}/E_j between successive echelons is a constant, then E_{tot} can be easily evaluated in terms of this ratio and the final maximum error E_n. Some values are given in Table I. We see that the total error is never more than twice the error added in the nth echelon, however many echelons there are, as long as the error ratio is at least two. However, if the error ratio approaches one, the total error mounts up rapidly.

Table I—Relative Total Maximum Error, E_{tot}/E_n

	E_{j+1}/E_j			
n	10	4	2	1
2	1.10	1.25	1.50	2.00
3	1.11	1.31	1.75	3.00
∞	1.11	1.33	2.00	∞

The above extreme case of perfect positive correlation would probably rarely be met in practice. Even though the error e_j passed on to a laboratory in echelon $(j + 1)$ has the effect of a systematic error within the calibrations performed by that laboratory until its next checkup with echelon j, or even over the course of many checkups, it is unlikely that the error e_{j+1} added by that laboratory in a particular case is appreciably correlated with e_j. In particular, if we restrict consideration of errors to uncorrelated deviations about mean values (i.e., to random errors), the variance of the total error, σ^2_{tot}, is given in terms of the variance σ_j^2 in the several echelons by the equation

$$\sigma^2_{tot} = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$$

Even if systematic as well as random errors are included, McNish and Cameron[3] pointed out that this square-type combination of errors is more realistic for a chain of calibrations than the simple sum. For simplicity we shall, however, give results for the square-type of combination in terms of the familiar notation σ. (See also Youden[8].)

If we assume that the error ratio σ_{j+1}/σ_j between successive echelons is a constant, then the standard deviation σ_{tot} is easily evaluated, and some relative values are given in Table II. Here we see that σ_{tot} is little more than the standard deviation in the last echelon, σ_n, however large the hierarchy is, unless σ_{j+1}/σ_j falls below two.

Cost Considerations

As indicated earlier, the optimum error ratio will be determined by minimizing the total cost of the entire hierarchy. For example, if the National Bureau of Standards were to require considerable basic research at great cost to improve its working standards, then to decrease the total error E_{tot} or σ_{tot} it would tend to be more economical to use more expensive instruments, methods, and personnel in the lower echelons. On the other hand, if the number of laboratories in lower echelons were very large, it would tend to be more economical to improve the single set of working standards at the Bureau, or the relatively few standards near the top echelon.

The costs to be considered are of two types: the cost of research and development that needs to be

Table II—Relative Total Standard Deviation, σ_{tot}/σ_n

	σ_{j+1}/σ_j			
n	10	4	2	1
2	1.005	1.03	1.12	1.41
3	1.005	1.03	1.15	1.73
∞	1.005	1.03	1.15	∞

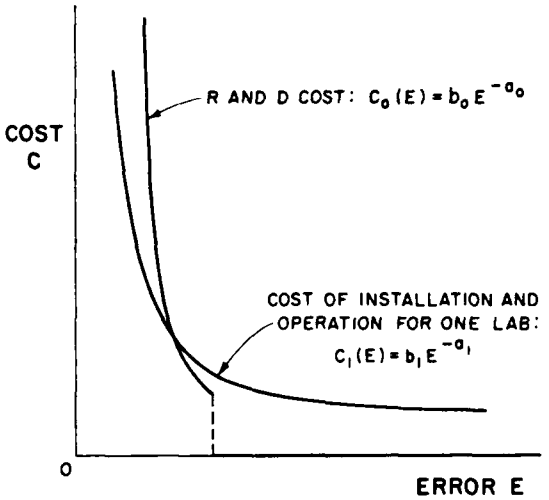


Figure 2—Illustrative Costs of Research and Development, $C_0(E)$, and of Installation and Operation, $C_1(E)$, Drawn for $a_0 = 2$ and $a_1 = 1$.

done only once, or even not at all if the system has already been developed, and costs of installation and operation for each laboratory. The latter costs need to be multiplied by the number of laboratories in each echelon. We assume that both types of costs are directly proportional to some power of the accuracy required; that is, inversely proportional to some power of the maximum or average error. Figure 2 shows functions of this type, with the research and development cost cut off at a certain error value to indicate that systems are already available for errors larger than that value. The exponents a_0 and a_1 are important; here the curves are drawn for $a_0 = 2$ and $a_1 = 1$, and in the general solution it is assumed that

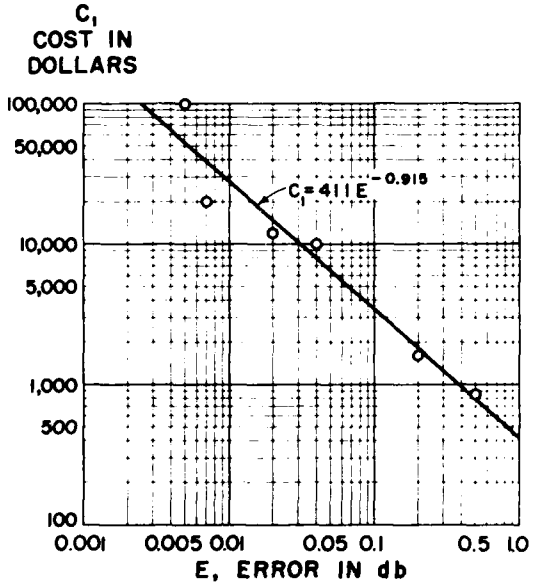


Figure 3—Estimated and Fitted Costs of Installation of Attenuator Calibration Systems.

Table III—Exponents a_1 for Four Calibration Systems

	Installation	Operation
Resistance	0.5	0.4
Voltage	0.5	0.2
Current	0.9	0.3
Power	1.0	0.3

$a_0 \geqq a_1$; i.e., that research and development cost, if present at all, rises more rapidly with accuracy than the costs of installation and operation.

In Fig. 3 are plotted the estimated costs of installing six attenuator calibration systems with different maximum errors E. These data were generously supplied to me by David H. Russell of the National Bureau of Standards. The best-fitting curve of the type in Fig. 2, fitted as the (least-squares) regression line of log C_1 on log E, is also shown; the exponent a_1, the slope, is about 0.9. From data kindly supplied to me by Frank D. Weaver and David Ramaley of the Bureau, I was also able to determine approximate exponents a_1 as shown in Table III.

The Optimization Problem

The essentials of the optimization problem can be best demonstrated on a simple example using the approximate costs of installing attenuator systems. Suppose there are only two echelons, with one laboratory in the first and eight in the second. Suppose also that the required maximum total error is 0.3 dB or about three percent, that each maximum error is taken as three times the standard deviation σ, and that errors combine by squares. Then the error equation is

$$\sigma_1^2 + \sigma_2^2 = \sigma_{tot}^2 = 1(\%)^2$$

(the coefficients three canceling). The cost of installing one system is, from Fig. 3, approximately $411/E_j$, where E_j is in dB, or

$$C_1(\sigma_j) = 1400/\sigma_j \text{ dollars } (\sigma_j \text{ in } \%),$$

The total cost equation therefore is

$$C_{tot}(\sigma_1, \sigma_2) = \frac{1400}{\sigma_1} + 8\,\frac{1400}{\sigma_2}$$

One can easily calculate this cost for various values of σ_1 and σ_2 whose squares add to one as required.

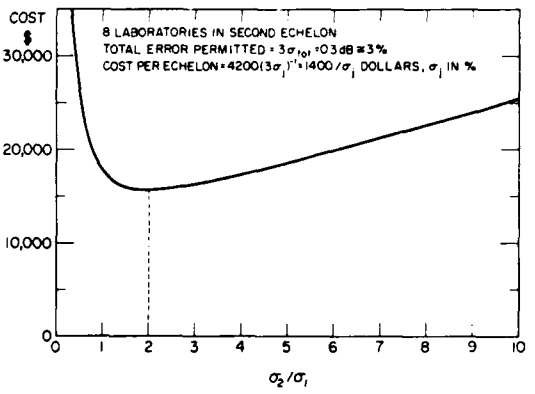


Figure 4—Theoretical Cost of Installing a Two-Echelon Hierarchy of Attenuator Calibration Systems.

In Fig. 4 the cost is graphed as a function of the error ratio σ_2/σ_1 (where σ_{tot} is kept equal to one percent always). We see that it is a minimum for $\sigma_2/\sigma_1 = 2$. This is confirmed by the general theory, which states that the optimum error ratio is the $(a_1 + 2)$th root of m_2, where m_2 is the number of laboratories in the second echelon, so that in this case we have the cube root of eight. The optimum errors themselves are easily found to be $\sigma_1 = 1/\sqrt{5} = 0.45\%$, $\sigma_2 = 2/\sqrt{5} = 0.89\%$. However, we see that the minimum of the curve is fairly flat; in fact the minimum cost of $15,650 for all nine laboratories would be exceeded only by five percent if the error ratio were as small as 1.3 or as large as 3.2.

If we had assumed that the errors combined linearly rather than by squares, then the optimum error ratio would have been the $(a_1 + 1)$th root of m_2, that is, the square root of eight, or 2.8 in this case. Thus, the practical difference between the two assumptions on error combinations is not large in this case.

On the other hand, we see that the value of the exponent a_1 could be important. If a_1 were very small relative to one, then the optimum error ratio would be close to m_2, or eight, if errors combine linearly. However, since it is more likely that squares of errors add, the optimum error ratio even for very small a_1 is more likely to be near the square root of m_2, or 2.8 here.

To consider the general optimization problem, let n be the number of echelons and m_j the number of laboratories in the jth echelon serviced by each laboratory in the $(j-1)$th echelon. For the case of linear combinations of errors the problem can be stated as follows:

Given a prescribed value of E_{tot}, determine E_1, E_2, . . . , E_n so as to minimize the total cost,

$$C(E_1, E_2, \ldots, E_n) = \sum_{j=1}^{n} [C_0(E_j) + m_1 m_2 \ldots m_j C_1(E_j)]$$

$$= \sum_{j=1}^{n} [b_0 E_j^{-a_0} + m_1 m_2 \ldots m_j b_1 E_j^{-a_1}],$$

subject to

$$E_1 + E_2 + \ldots + E_n = E_{tot},$$

where $a_0 \geq a_1 > 0$ and b_0 may be 0 for E_j sufficiently large.

The solution to this problem can be obtained by applying differential calculus (using the "Method of Lagrange Multipliers"). The optimum error ratio between each pair of echelons is given by

$$E_{j+1}/E_j = m_{j+1}^{1/(a+1)} \qquad (j = 1, 2, \ldots, n-1),$$

where $0 < a \leq a_1$, and, in the particular case that research and development are unnecessary, $a = a_1$. Table IV gives some values of these optimum ratios for a wide range of values of m_2 and three values of a. For the defense hierarchies shown in Fig. 1 the m_j's vary from 2 to 23. Only for the laboratories in the echelon reporting to the National Bureau of Standards does m_j rise to as high as 125, and that figure is for a very common quantity like resistance when all the industrial laboratories as well as the military laboratories are included.

The solution to the corresponding problem with errors being combined by squares is given by the optimum error ratio

$$\sigma_{j+1}/\sigma_j = (m_{j+1})^{1/(a+2)} \qquad (j = 1, 2, \ldots, n-1),$$

where $0 < a \leq a_1$ and $a = a_1$ in the particular case that research and development are unnecessary. Table V gives some values of these optimum ratios, all of which are smaller than the corresponding values in Table IV. Since errors are more likely to combine as squares than linearly, Table V should be considered as more appropriate than Table IV. From our cost estimates on electronic quantities, the column $a = 0.3$ seems to apply approximately to operating costs and the column $a = 0.8$ to installation costs. Thus for the m_j values occurring in the military calibration hierarchies, the optimum error ratios range from about 1.3 to 4.

Concluding Remarks

We have considered the errors occurring in a hierarchy of calibrations from two points of view, how the errors probably accumulate and how they *should* accumulate, or be allocated, from the point of view of minimizing the total cost. We have seen that from either point of view the errors add up rather slowly and that the optimum ratio of errors from one echelon to the next should be relatively small in most cases, perhaps about 1.3 to 5. When the size and a cost exponent of a particular calibration hierarchy are known, the optimum error ratio can be calculated from the formulas given. These results are valid only to the extent that the assumptions, the cost functions of Fig. 2 in particular, are realistic.

Acknowledgments

In addition to the invaluable aid of David H. Russell, Frank D. Weaver, and David Ramaley noted at particular points, I am indebted to several other staff members of the National Bureau of Standards for help in the evolution of the work described: to Churchill Eisenhart for calling attention to the problem, to A. G. McNish concerning the combination of errors, to Harvey W. Lance, Wilbert F. Snyder, and Thomas L. Zapf for calibration information and suggestions, and to Charles L. Bragaw for some of the historical introduction.

Table IV—Optimum Error Ratios if $E_{tot} = E_1 + E_2 + \ldots + E_n$

	$E_{j+1}/E_j = m_{j+1}^{1/(a+1)}$		
m_{j+1}	a = 0.3	a = 0.8	a = 2
2	1.7	1.5	1.3
8	5.0	3.2	2.0
32	14.4	6.9	3.2
128	41.6	14.8	5.0

Table V—Optimum Error Ratios if $\sigma^2_{tot} = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$

	$\sigma_{j+1}/\sigma_1 = m_{j+1}^{1/(a+2)}$		
m_{j+1}	a = 0.3	a = 0.8	a = 2
2	1.4	1.3	1.2
8	2.5	2.1	1.7
32	4.5	3.4	2.4
128	8.2	5.7	3.4

References

1. Crow, Edwin L., "An Analysis of the Accumulated Error in a Hierarchy of Calibrations", *IRE Transactions on Instrumentation, I-9,* 105-114 (Sept. 1960).
2. Eisenhart, Churchill, "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems", *Jour. of Research of the National Bureau of Standards, 67C,* 161-187 (Apr.-Jun. 1963).
3. McNish, A. G., and Cameron, J. M., "Propagation of Error in a Chain of Standards", *IRE Transactions on Instrumentation, I-9,* 101-104 (Sept. 1960).
4. National Bureau of Standards, *Technical News Bulletin, 39,* 57 (Apr. 1955).
5. Perry, John, *The Story of Standards,* Funk & Wagnalls Company, New York, 1955.
6. Smith, Ralph W., *The Federal Basis for Weights and Measures,* National Bureau of Standards Circular 593, U.S. Government Printing Office, Washington, D.C., June 5, 1958. A more recent reference is: Judson, Lewis V., *Weights and Measures Standards of the United States: A Brief History,* National Bureau of Standards Miscellaneous Publication 247. U.S. Government Printing Office, Washington, D.C., October 1963.
7. Woods, W. Max, and Zehna, Peter W., "Cumulative Effect of Calibration Errors", *Ind. Qual. Control, 22,* 411-412 (Feb. 1966).
8. Youden, W. J., "Uncertainties in Calibration", *IRE Transactions on Instrumentation, I-11,* 133-138 (Dec. 1962).

CONFIDENCE AND TOLERANCE INTERVALS FOR THE NORMAL DISTRIBUTION*

FRANK PROSCHAN
National Bureau of Standards†

Confidence and tolerance intervals for the normal distribution are presented for the various cases of known and unknown mean and standard deviation. Practical illustration and interpretation of these intervals are given. Tables are presented permitting a comparison among the intervals. Finally the relationship between the two types of intervals is described.

1. *Introduction.* Discussions of the theory of errors will sometimes state that the mean plus or minus the probable error will include 50% of future observations (assumed normally distributed). This, of course, is true only if the mean and the probable error of the population itself are used. Unfortunately, in most practical problems one or both of these may not be known. Experimenters who use the *sample* mean plus or minus the sample probable error with the expectation that this interval will contain 50% of future observations may be seriously deluding themselves.

However it *is* possible to construct intervals of the type $\bar{x} \pm ks$ ($\bar{x}$ = sample mean, s = sample standard deviation) which will, on the average, include 50% of the population. From this one is led to a more general consideration of such intervals, and to the uses to which they can be put.

All populations discussed in this paper are normal unless otherwise specified. Let μ, σ refer to the population mean and standard deviation respectively.

Any one of four possible situations may exist: (a) μ, σ both known; (b) μ unknown, σ known; (c) μ known, σ unknown; (d) μ, σ both unknown.

Let m represent either μ or $\bar{x}$; let $s.d.$ represent either σ or s. Then two important types of assertions may be made about intervals of the form

$$m \pm k\ s.d. \tag{1}$$

A. *Confidence interval.* The probability is γ that the interval (1) contains the population mean (or, alternatively, the second sample mean).

* Presented at the annual meeting of the American Statistical Association, Boston, December 1951.
† Now at Syracuse Electric Products, Inc., Hicksville, N. Y.

B. *Tolerance interval.* In a large series of repeated samples the proportion of the population contained in (1) is

(B1) a, on the average

(B2) P or more, γ of the time.

In this paper, a comparison is made among the values of k appropriate to the respective cases obtained from various combinations of A and B with (a), (b), (c), and (d). Practical illustrations and interpretations are given of these cases.

In addition, details are given of a proof of a result by Wilks (1941) for the case B1. These details are given because they are suggestive of a general method applicable in such problems. Also, tables are presented of values of k for a certain class of confidence and tolerance intervals.

Finally, the relationship between confidence intervals and tolerance intervals is discussed.

2. *Definition of symbols.* For convenience, the definitions of symbols are brought together in this section.

μ = population mean

σ = population standard deviation

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ mean of a sample of } n \text{ observations,}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}, \text{ sample standard deviation}$$

$$m = \mu \quad \text{or} \quad \bar{x}$$

$$s.d. = \sigma \quad \text{or} \quad s$$

p = proportion of the population contained in $m \pm k$ $s.d.$ where k = constant

Given a normal distribution with $\mu = 0$, $\sigma = 1$. Then

L_α = normal curve deviate which is exceeded in absolute value with probability α

$t_{\alpha,n-1}$ = Student-t value for $n-1$ degrees of freedom which is exceeded in absolute value with probability α.

$\chi^2_{\alpha,n-1}$ = Chi-square value with $n-1$ degrees of freedom which is exceeded with probability α.

3. *Confidence intervals.* A chemist makes n determinations of the iron content, μ, of a solution. What interval shall he select so that he can

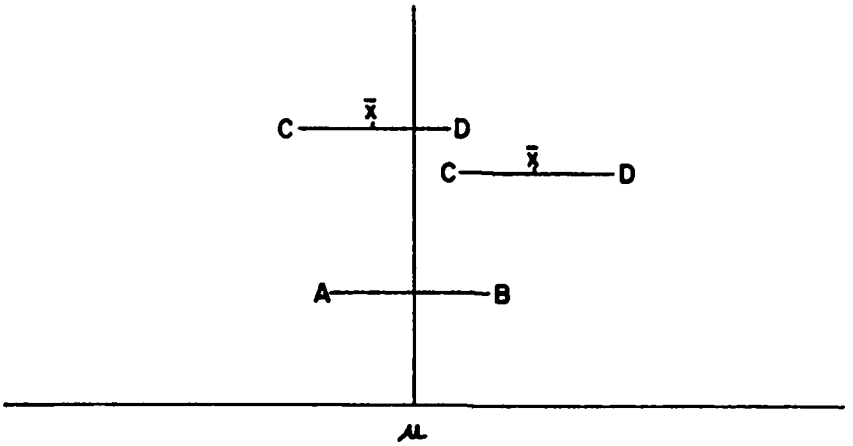
assert with 50% confidence that μ lies within that interval? The distribution of observations is normal with mean μ.

3.1 *For the population mean, standard deviation known.* First consider the case where the chemist knows σ. (The determination is of a routine type, for which a great many sets of previous observations are available, from which σ is calculated.) In this case

$$\bar{x} \pm \frac{.6745}{\sqrt{n}}\sigma$$

will contain the "true" value (population mean) 50% of the time.

This may be seen from the following diagram:



Lay off $AB: \mu \pm (.6745/\sqrt{n})\sigma$, and $CD: \bar{x} \pm (.6745/\sqrt{n})\sigma$. Notice that when $\bar{x}$ lies in AB, μ must of necessity lie in CD; and when $\bar{x}$ does not lie in AB, μ must lie outside of CD. But since $\bar{x}$ *is* normally distributed with mean μ, standard deviation $(\sigma/\sqrt{n})$ the probability is .50 that $\bar{x}$ will lie in AB. Hence the probability is .50 that CD contains μ.

Values of $k_1 = .6745/\sqrt{n}$ for $n = 2(1)$ 30, 40, 60, 120, ∞ are presented in Table 1.

To generalize, when the confidence coefficient is γ, the confidence interval for μ is

$$\bar{x} \pm \frac{L_{1-\gamma}}{\sqrt{n}}\sigma$$

3.2 *For the population mean, standard deviation unknown.* Consider the case where the only information about σ is in the present sample.

Then the interval

$$\bar{x} \pm \frac{t_{.50,n-1}}{\sqrt{n}} s$$

will contain μ, 50% of the time. The proof is similar to that of Section 3.1. Values of $k_2 = t_{.50,n-1}/\sqrt{n}$ for $n = 2(1)$ 30, 40, 60, 120, ∞ are presented in Table 1. Comparison of k_1 and k_2 shows $k_2 > k_1$, but as $n \to \infty$, $k_2 \to k_1$.

In general, when the confidence coefficient is γ, the confidence interval is

$$\bar{x} - \frac{t_{1-\gamma,n-1}}{\sqrt{n}} s \leqq \mu \leqq \bar{x} + \frac{t_{1-\gamma,n-1}}{\sqrt{n}} s.$$

3.3 *"Confidence interval" for second sample mean.* Suppose the chemist who made the iron determinations wishes to set up a confidence interval, not for μ, but for the mean, $\bar{x}_2$, of a second sample of n_2 observations. Such an interval might be called more appropriately a prediction interval, since the term "confidence interval" generally refers to population parameters.

Let us call the mean of the first sample $\bar{x}_1$, and its size n_1. Since the statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_1 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is distributed as the Student-t ratio, it follows that the interval

$$\bar{x}_1 \pm t_{.50,n_1-1} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, s_1 \tag{2}$$

will constitute a 50% prediction interval for $\bar{x}_2$ [1].

What does this mean? It simply means that if pairs of samples of size n_1 and n_2 respectively, with means $\bar{x}_{1i}$ and $\bar{x}_{2i}$ $(i = 1, 2, \cdots)$ respectively are drawn repeatedly, then for 50% of these pairs $\bar{x}_{2i}$ will lie in

$$\bar{x}_{1i} \pm t_{.50,n_1-1} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, s_{1i}.$$

It does *not* mean that if *one* first sample of size n_1 with mean $\bar{x}_1$ is drawn, to be followed by the drawing of a great many "second" samples of

TABLE 1

FACTORS FOR 50% CONFIDENCE INTERVALS

For	μ	μ	$\bar{x}_2$
σ known (K) or unknown (U)	K	U	U
Form of interval	$\bar{x} \pm k_1\sigma$	$\bar{x} \pm k_2 s$	$\bar{x}_1 \pm k_3 s$
n	k_1	k_2	k_3
2	.477	.707	1.000
3	.389	.471	.666
4	.337	.382	.541
5	.302	.331	.469
6	.275	.297	.420
7	.255	.271	.384
8	.238	.251	.356
9	.225	.235	.333
10	.213	.222	.314
11	.203	.211	.299
12	.195	.201	.285
13	.187	.193	.273
14	.180	.185	.262
15	.174	.179	.253
16	.169	.173	.244
17	.164	.167	.237
18	.159	.162	.230
19	.155	.158	.223
20	.151	.154	.218
21	.147	.150	.212
22	.144	.146	.207
23	.141	.143	.202
24	.138	.140	.198
25	.135	.137	.194
26	.132	.134	.190
27	.130	.132	.186
28	.127	.129	.183
29	.125	.127	.179
30	.123	.125	.176
40	.107	.108	.152
60	.087	.088	.124
120	.062	.062	.087
∞	0	0	0
For discussion see Section	3.1	3.2	3.2

size n_2 with means $\bar{x}_{2i}$ $(i=1, 2 \cdots)$ then for 50% of the "second" samples $\bar{x}_{2i}$ will lie in (2).

When $n_2=n_1=n$ the coefficient of s_1 in (2) becomes

$$k_3 = t_{.50,n-1}\sqrt{\frac{2}{n}}$$

Values of k_3 for $n=2$ (1) 30, 40, 60, 120, ∞ are given in Table 1 for purposes of comparison. Note that $k_3=\sqrt{2}k_2$ simply.

In general, if the "confidence" coefficient is to be γ for $\bar{x}_2$, then the interval to be used is

$$\bar{x}_1 \pm t_{1-\gamma,n_1-1}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\,s_1.$$

4. *Tolerance intervals.* In Section 3 an interval of type (1) was formed to contain the population mean (with a certain confidence). Suppose now, we are interested in setting up an interval of type (1) which will contain a certain *proportion p of the population.* Such an interval is known as a tolerance interval.

If either μ or σ is unknown, then the interval (1), containing $\bar{x}$ or s, is random. Hence the proportion p contained in (1) will be a random variable.

4.1 *Average value of p.* In Section 4.1 we determine k so that on the average the proportion p is equal to a, a constant. In Section 4.2 we determine k so that in a large series of samples from normal universes a certain proportion γ of the intervals (1) will include a proportion p or more of the universe.

4.1.1 *Population mean and standard deviation known.* In this case

$$\mu \pm k\sigma \tag{3}$$

may be used as a tolerance interval. The proportion p contained in (3) is constant, and the appropriate value for specified p may be obtained from a table of normal areas. Thus for $p=.50$, $k_4=.6745$ (listed in Table 2 for purposes of comparison).

4.1.2 *Population mean and standard deviation unknown.* Unfortunately, in most practical problems μ and σ are not known. Hence $\bar{x}$ and s must be used. How shall we determine k so that the average p contained in $\bar{x}_i \pm ks_i$ $(i=1, 2, \cdots)$ will be a?

Wilks [8] gave a solution without presenting the details of the proof. (For an independent derivation see Appendix.) The solution states that the tolerance limits which will include, on the average, a proportion a of the normal universe are

$$\bar{x} \pm t_{1-a,n-1}\sqrt{\frac{n+1}{n}}\,s. \tag{4}$$

Values of $k_{a,a}=t_{1-a,n-1}\sqrt{n+1/n}$ for $n=2$ (1) 30, 40, 60, 120, ∞ and for $a=.50, .75, .90, .95, .99, .999$ are given in Table 3. This table should be of use both to the experimenter and to the quality controller; it supplements the values of k given in [3]. In addition, for purposes of comparison, Table 2 gives values of $k_{5,.50}$ for $n=2$ (1) 30, 40, 60, 120, ∞.

An example of the use of Table 3 is given:

Example: A quality control engineer measures the voltages of a random sample of 30 batteries from his production line. From the sample mean voltage $\bar{x}=7.52$ and the sample standard deviation of voltages $s=.90$, he wishes to estimate tolerance limits that will, on the average, contain 95% of the population of batteries. Assuming the distribution of battery voltages to be normal, what shall these tolerance limits be?

The tolerance limits will be of the form $\bar{x}\pm k_{5,.95}s$. Entering Table 3 with $n=30$, he finds $k_{5,.95}=2.079$. Hence the tolerance limits are:

$$7.52 \pm 2.079(.90) = 7.52 \pm 1.87 = 5.65 \text{ to } 9.39.$$

Notice that $k_{5,.95}=2.079$ is larger than the value 1.96 that would be used if μ and σ were known.

One sided tolerance limits. Suppose now the problem is to find the value of k_a' such that, on the average, the proportion of the normal population *less than* $\bar{x}+k_a's$ is a specified value a. By the same procedure as in the proof for the two sided case (Appendix) it may be shown that

$$k_a' = k_{5,2a-1}. \tag{5}$$

A similar result holds if the proportion of the normal population *greater than* $\bar{x}-k_a's$ is to be a specified value a, on the average.

Example: A pilot run of 40 electron tubes is made. For each tube the plate current in milliamperes, x, under normal operating voltages, is measured; for the sample $\bar{x}=12.25$, $s=.68$. From past experience with similar tubes, it is known that x is normally distributed. What procedure shall be followed to determine the value of L such that 99% of the population of tubes will, on the average, have a plate voltage less than L?

We may write

$$L = \bar{x} + k_{.99}'s.$$

Then according to (5)

$$k_{.99}' = k_{5,2(.99)-1} = k_{5,.98}.$$

Table 3 furnishes $k_{5,.98} = 2.455$. Hence

$$L = 12.25 + 2.455(.68) = 13.92.$$

4.1.3 *Population mean unknown, standard deviation known.* In this case an interval of the form

$$\bar{x} \pm k_6\sigma \tag{6}$$

must be used. Using the same method as in the proof given in the Appendix, the following result may be derived:

If the expected value $E(p)$ of the proportion p of the normal universe contained in (6) is to be a, then

$$k_6 = \sqrt{\frac{n+1}{n}}\, L_{1-a}.$$

For purposes of comparison, k_6 is given in Table 2 for $a = .50$ and for $n = 2(1)$ 30, 40, 60, 120, ∞.

4.1.4 *Population mean known, standard deviation unknown.* In this case the interval

$$\mu \pm k_7 s \tag{7}$$

must be used. Again using the same method as in the proof of the Appendix, the appropriate value for k_7 to include, on the average, a is given by

$$k_7 = t_{1-a,n-1}.$$

For purposes of comparison, values of k_7 are given in Table 2, for $a = .50$ and $n = 2$ (1) 30, 40, 60, 120, ∞.

4.2 *Confidence statement about tolerance interval.* A number of papers have been written on the problem of confidence statements for tolerance intervals [2], [3], [6], [7], [8], [9]. The problem may be illustrated as follows:

4.2.1 *Population mean and standard deviation unknown.* Suppose the battery engineer mentioned in Section 4.1.2 asked the following question: What value of k shall I take so that I can be 95% confident that $\bar{x} \pm ks$ will include at least 80% of my population of batteries?

Bowker [3, pp. 102–107] gives extensive tables of k such that "in

TABLE 2

FACTORS FOR TOLERANCE INTERVALS

that will include, on the average, 50% of the population. or that will include at least 50% of the population 50% of the time.	√	√	√	√	√	√
μ known (K) or unknown (U)	K	U	U	K	K	U
σ known (K) or unknown (U)	K	U	K	U	U	K
Form of interval	$\mu \pm k_4\sigma$	$\bar{x} \pm k_5 s$	$\bar{x} \pm k_6\sigma$	$\mu \pm k_7 s$	$\mu \pm k_8 s$	$\bar{x} \pm k_9\sigma$
n	k_4	$k_{5..50}$	k_6	k_7	k_8	k_9
2	.674	1.225	.826	1.000	1.000	754
3	.674	.942	.779	.816	.810	.727
4	.674	.855	.754	.765	.759	.714
5	.674	.812	.739	.741	.736	.706
6	.674	.785	.729	.727	.723	.700
7	.674	.768	.721	.718	.714	.697
8	.674	.754	.715	.711	.708	.694
9	.674	.744	.711	.706	.704	.692
10	.674	.737	.707	.703	.701	.690
11	.674	.731	.704	.700	.698	.688
12	.674	.725	.702	.697	.698	.687
13	.674	.721	.700	.695	.694	.686
14	.674	.718	.698	.694	.692	.686
15	.674	.715	.697	.692	.691	.685
16	.674	.712	.695	.691	.690	.684
17	.674	.710	.694	.690	.689	.684
18	.674	.708	.693	.689	.688	.683
19	.674	.706	.692	.688	.687	.683
20	.674	.705	.691	.688	.687	.682
21	.674	.703	.690	.687	.686	.682
22	.674	.701	.690	.686	.685	.681
23	.674	.701	.689	.686	.685	.681
24	.674	.699	.688	.685	.684	.681
25	.674	.699	.688	.685	.684	.681
26	.674	.697	.687	.684	.684	.680
27	.674	.697	.687	.689	.683	.680

TABLE 2 (*cont.*)

n	k_4	$k_{5,.50}$	k_6	k_7	k_8	k_9
28	.674	.696	.686	.684	.683	.680
29	.674	.695	.686	.683	.683	.680
30	.674	.694	.686	.683	.682	.680
40	.674	.689	.683	.681	.680	.678
60	.674	.685	.680	.679	.678	.677
120	.674	.680	.677	.677	.676	.676
∞	.674	.674	.674	.674	.674	.674
For discussion see Section	4.1.1	4.1.2	4.1.3	4.1.4	4.2.2	4.2.3

a large series of samples for normal universes a certain proportion γ of the intervals $\bar{x} \pm ks$ will include P or more of the universe; γ is called the "confidence coefficient" since it is a measure of the confidence with which we may assert that a given tolerance range includes at least P of the universe." In these tables $\gamma = .75, .90, .95, .99, .999$.

4.2.2 *Population mean known, standard deviation unknown.* Consider the case where μ is known and δ unknown. Then an interval of the form

$$\mu \pm k_8 s \tag{8}$$

can be set up to include at least a proportion P of the population with confidence γ as follows:

Let us take specific values of $P = .80$ and $\gamma = .95$ for illustrative purposes. We note first that p is monotonic increasing with s (and with s^2). Hence when s^2 takes on a value exceeded 95% of the time (call it $s_{.95}^2$), p will take on a value exceeded 95% of the time. But

$$s_{.95}^2 = \frac{\chi_{.95,n-1}^2}{n-1}\sigma^2.$$

Then the appropriate value of k_8 is

$$k_8 = L_{.20}/\sqrt{\chi_{.95,n-1}^2/(n-1)}.$$

Values of k_8 for $p = \gamma = .50$ for $n = 2$ (1) 30, 40, 60, 120, ∞. are given in Table 2, for purposes of comparison.

For general P, γ

$$k_8 = L_{1-P}/\sqrt{\chi^2_{\gamma,n-1}/(n-1)}.$$

4.2.3 *Population mean unknown, standard deviation known.* In this case, an interval of the type

$$\bar{x} \pm k_9\sigma \tag{9}$$

must be used. Let us solve for k_9 when $P = .80$, $\gamma = .95$ to illustrate the reasoning.

We first note that 95% of the $\bar{x}$'s lie in the interval $\mu \pm (L_{.05}/\sqrt{n})\sigma$ that is, 95% of the $\bar{x} \pm k_9\sigma$ intervals have their centers inside the interval $\mu \pm (L_{.05}/\sqrt{n})\sigma$. Now we find k_9 such that the normal curve area between $\mu + (L_{.05}/\sqrt{n})\sigma - k_9\sigma$ and $\mu + (L_{.05}/\sqrt{n})\sigma + k_9\sigma$ is .80. Then 95% of the $\bar{x} \pm k_9\sigma$ intervals will contain $p \geqq .80$ (namely those intervals for which $\bar{x}$ lies in $\mu \pm (L_{.05}/\sqrt{n})\sigma$).

It follows that the interval (9) will contain a proportion .80 or more of the population, .95 of the time.

Values of k_9 for $P = \gamma = .50$ are given in Table 2 for $n = 2$ (1) 30, 40, 60, 120, ∞. For general P, γ, k_9 is the value such that the normal curve area between $\mu + (L_{1-\gamma}/\sqrt{n})\sigma - k_9\sigma$ and $\mu + (L_{1-\gamma}/\sqrt{n})\sigma + k_9\sigma$ is P.

5. *Relationship between confidence intervals and tolerance intervals.* There is a very interesting relationship between confidence intervals and tolerance intervals that may be illustrated by the following example:

Suppose, as in Section 3.3, we wanted to find a prediction (or "confidence") interval for the mean of a second sample. But now let $n_2 = 1$. In other words, we will now be finding a confidence interval for a single future observation. According to the result in Section 3.3. our answer is

$$\bar{x}_1 \pm t_{1-a,n_1-1}\sqrt{\frac{1}{n_1} + \frac{1}{1}}\, s_1 = \bar{x}_1 \pm t_{1-a,n_1-1}\sqrt{\frac{n_1+1}{n_1}}\, s_1 \tag{10}$$

where a is the confidence coefficient.

What does this mean? One way of looking at it is that if repeatedly a sample of size n_1 is first drawn and then a second sample of one item is drawn, then a proportion a of the time the single item will lie in the interval (10). But a little thought shows that this is exactly equivalent to stating that in repeated samples of size n_1, the average proportion, p, of the population contained in (10) is a. In other words, confidence limits with confidence coefficient a for a second sample of size one are identical with tolerance limits that will include a proportion a on the average. This is confirmed by the fact that (10) is the same as (4) (except for the subscript 1).

The above is an illustration of a theorem by Paulson [5]:

"If confidence limits $U_1(x_1, \ldots, x_n)$ and $U_2(x_1, \ldots, x_n)$ on a probability level $= \alpha_0$ are determined for g, a function of a future sample of

TABLE 3

FACTORS, $k_{5,a}$ FOR TOLERANCE INTERVALS SUCH THAT $\bar{x} \pm k_{5,a}\, s$ WILL INCLUDE A PROPORTION a OF THE POPULATION, ON THE AVERAGE

Sample size, n	$k_{5,.50}$	$k_{5,.75}$	$k_{5,.90}$	$k_{5,.95}$	$k_{5,.98}$	$k_{5,.99}$	$k_{5,.999}$
2	1.225	2.957	7.733	15.562	38.973	77.964	779.699
3	.942	1.852	3.372	4.969	8.042	11.460	36.486
4	.855	1.591	2.631	3.558	5.077	6.530	14.469
5	.812	1.473	2.335	3.041	4.105	5.043	9.432
6	.785	1.405	2.176	2.777	3.635	4.355	7.409
7	.768	1.361	2.077	2.616	3.360	3.963	6.370
8	.754	1.330	2.010	2.508	3.180	3.711	5.733
9	.744	1.307	1.961	2.431	3.053	3.536	5.314
10	.737	1.290	1.922	2.372	2.959	3.409	5.014
11	.731	1.276	1.893	2.327	2.887	3.310	4.791
12	.725	1.264	1.869	2.291	2.829	3.233	4.618
13	.721	1.255	1.849	2.261	2.782	3.170	4.481
14	.718	1.246	1.833	2.236	2.743	3.118	4.369
15	.715	1.239	1.819	2.215	2.710	3.075	4.276
16	.712	1.234	1.807	2.197	2.682	3.038	4.198
17	.710	1.228	1.797	2.181	2.658	3.006	4.131
18	.708	1.224	1.788	2.168	2.637	2.977	4.074
19	.706	1.220	1.779	2.156	2.618	2.953	4.024
20	.705	1.216	1.772	2.145	2.602	2.932	3.979
21	.703	1.213	1.766	2.135	2.587	2.912	3.941
22	.701	1.210	1.760	2.127	2.575	2.895	3.905
23	.701	1.207	1.754	2.119	2.562	2.880	3.874
24	.699	1.205	1.749	2.112	2.552	2.865	3.845
25	.699	1.202	1.745	2.105	2.541	2.852	3.819
26	.697	1.200	1.741	2.099	2.532	2.840	3.796
27	.697	1.198	1.737	2.094	2.524	2.830	3.775
28	.696	1.197	1.733	2.088	2.517	2.820	3.755
29	.695	1.196	1.730	2.083	2.509	2.810	3.737
30	.694	1.193	1.727	2.079	2.503	2.802	3.719
40	.689	1.182	1.706	2.047	2.455	2.741	3.602
60	.685	1.171	1.686	2.017	2.411	2.684	3.492
120	.680	1.161	1.665	1.988	2.368	2.628	3.388
∞	.674	1.150	1.645	1.960	2.326	2.576	3.291

See Section 4.1.2 for a discussion of this case.

k observations, with distribution $\Psi(g)$, and $p=\int_{U_1}^{U_2}\psi(g)dg$, then $E(p)=\alpha_0$."

In the illustration of this section, g corresponds to the value of the single future observation and $k=1$. Similarly we can check the results of Sections 4.1.3 and 4.1.4 by the use of Paulson's theorem.

APPENDIX

Mathematical proof of Wilks' result. The details of the derivation (independently obtained by I. R. Savage of the Statistical Engineering Laboratory, National Bureau of Standards) of the result of Section 4.1.2 are given, since the method is a suggestive one.

The problem is to determine k so that the average p contained in $\bar{x}_i \pm ks_i$ $(i=1, 2, \ldots)$ will be a. By an appropriate linear transformation, the problem may be reduced to that of finding

$$E(p) = C_1 \int_0^\infty \int_{-\infty}^\infty \int_{\bar{x}-ks}^{\bar{x}+ks} e^{-\frac{1}{2}t^2} dt s^{n-2} e^{-\frac{1}{2}[n\bar{x}^2+(n-1)s^2]} d\bar{x}ds$$

where C_1 is a constant free of k. In the following, C_i = constant free of k.

The conditions for differentiating under the integral hold. Hence we have

$$\frac{\partial E}{\partial k} = C_1 \int_0^\infty \int_{-\infty}^\infty \left[s e^{-\frac{1}{2}(\bar{x}+ks)^2} + s e^{-\frac{1}{2}(\bar{x}-ks)^2} \right] s^{n-2} e^{-\frac{1}{2}[n\bar{x}^2+(n-1)s^2]} d\bar{x}ds$$

$$= C_1 \int_0^\infty \int_{-\infty}^\infty e^{-\frac{1}{2}[(\sqrt{n+1}\,\bar{x}+(ks/\sqrt{n+1}))^2+(n-1+k^2(n/n+1))s^2]} s^{n-1} d\bar{x}ds$$

$$+ C_1 \int_0^\infty \int_{-\infty}^\infty e^{-\frac{1}{2}[(\sqrt{n+1}\,\bar{x}-(ks/\sqrt{n+1}))^2+(n-1+k^2(n/n+1))s^2]} s^{n-1} d\bar{x}ds,$$

$$\frac{\partial E}{\partial k} = C_1 \int_0^\infty \int_{-\infty}^\infty e^{-\frac{1}{2}u^2} \frac{du}{\sqrt{n+1}} s^{n-1} e^{-\frac{1}{2}(n-1+k^2(n/n+1))s^2} duds$$

$$+ C_1 \int_0^\infty \int_{-\infty}^\infty e^{-\frac{1}{2}u^2} \frac{du}{\sqrt{n+1}} s^{n-1} e^{-\frac{1}{2}(n-1+k^2(n/n+1))s^2} duds,$$

$$\frac{\partial E}{\partial k} = C_2 \int_0^\infty s^{n-1} e^{-\frac{1}{2}(n-1+k^2(n/n+1))s^2} ds.$$

Let

$$y = \tfrac{1}{2}\left(n - 1 + k^2 \frac{n}{n+1}\right) s^2.$$

Then

$$\frac{\delta E}{\delta k} = C_2 \int_0^\infty 2^{n/2-1} y^{n/2-1} e^{-y} \Big/ \left(n - 1 + k^2 \frac{n}{n+1}\right)^{n/2} dy$$

$$= C_3 \frac{1}{\left(n - 1 + k^2 \frac{n}{n+1}\right)^{n/2}}.$$

Hence

$$E(p) = C_3 \int_{-k}^{k} \frac{dk}{\left(n - 1 + k^2 \frac{n}{n+1}\right)^{n/2}}.$$

Now let

$$t = k\sqrt{\frac{n}{n+1}},$$

so that

$$E(p) = C_4 \int_{-t}^{t} \frac{dt}{(n - 1 + t^2)^{n/2}}$$

$$= C_5 \int_{-t}^{t} dt/(1 + t^2)/(n - 1))^{n/2}.$$

But the integrand is the well known Student-t density function. Now when $k = \infty$, $E(p) = 1$. Hence C_5 must be identical with the constant of the Student-t distribution. Therefore the result of Section 4.1.2 follows:

$$k = t_{1-a,n-1}\sqrt{\frac{n+1}{n}}.$$

[1] Baker, G. A., "The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample," *Annals of Mathematical Statistics*, 6 (1935), 197–201.

[2] Bowker, A. H., "Computation of factors for tolerance limits on a normal distribution when the sample is large," *Annals of Mathematical Statistics*, 17 (1946), 238–40.

[3] Bowker, A. H., "Tolerance limits for normal distributions," Chapter 2 of Statistical Research Group, Columbia University, *Techniques of Statistical Analysis*. New York: McGraw-Hill (1947), 95–110.
[4] Mood, A. M., *Introduction to the Theory of Statistics*. New York: McGraw-Hill (1950), Chapter 11, 220–44.
[5] Paulson, E., "A note on tolerance limits," *Annals of Mathematical Statistics*, 14 (1943), 90–3.
[6] Wald, A., "Setting of tolerance limits when the sample is large," *Annals of Mathematical Statistics*, 13 (1942), 389–99.
[7] Wald, A., and Wolfowitz, J., "Tolerance limits for a normal distribution," *Annals of Mathematical Statistics*, 17 (1946), 208–15.
[8] Wilks, S. S., "Determination of sample sizes for setting tolerance limits," *Annals of Mathematical Statistics*, 12 (1941), 91–6.
[9] Wallis, W. Allen, "Tolerance intervals for linear regression," in *Second Berkeley Symposium on Mathematical Statistics and Probability*, edited by Jerzy Neyman. Berkeley: University of California Press (1951), 43–51.

Reprinted from the Journal of the American Statistical Association, Vol. 48, 550-564, 1953.

THE RELATION BETWEEN CONFIDENCE INTERVALS AND TESTS OF SIGNIFICANCE

— A Teaching Aid —

by Mary G. Natrella, *National Bureau of Standards*

1. Introduction.

The advertising sheet for a recent revision of a classical text book on statistical methods states "The author has diverted emphasis from tests of significance to point and interval estimates." This author is not alone. Many statistical consultants, analyzing an experiment for the purpose of testing a statistical hypothesis, e.g., in comparing means of normal populations, find that they prefer to present results in terms of the appropriate confidence interval.

It must be noted of course that not every statistical test can be put in the form of a confidence interval. Kendall, [5], for example, speaks of two broad classes of statistical tests, "those which give a direct test of a given value of a parent parameter, and those which do not." Berkson [2] also distinguishes these two classes of tests in discussing tests of normality and says "I suggest tentatively that the two classes I have in mind can be differentiated as (1) those which in principle can be alternatively stated in terms of an estimate and its confidence interval and (2) those which cannot be so stated." It is this first class of tests which will be discussed in this paper. Tests such as tests of normality, tests of goodness-of-fit, and tests of randomness fall into the second class.

When the results of a statistical test can alternatively be stated in terms of a confidence interval for a parameter, is there any reason to prefer the confidence interval statement? An early indication of dissatisfaction with the logic of tests of significance as experimental evidence is given in another paper by Berkson [3]. He stresses the point that experimenters are not typically engaged in disproving things, but are looking for evidence for affirmative conclusions, and that after rejecting the null hypothesis, they will then look for a reasonable hypothesis to accept. The relation between confidence intervals and tests of significance is mentioned only briefly by most textbooks, and ordinarily no insight is given as to which conclusion might be more appropriate. (A notable exception is Wallis and Roberts [7].)

In the present note, we draw attention to how these two approaches are related and how they differ. One reason for preferring to present a confidence interval statement (where possible) is that the confidence interval, by its width, tells more about the reliance that can be placed on the results of the experiment than does a YES-NO test of significance. Of course, a test of significance, when accompanied by its appropriate Operating Characteristic curve, provides much the same kind of information as does a confidence interval. In practice, however, the associated O.C. curve is often ignored by and may be unknown to the experimenter. We feel that the experimenter himself finds the confidence interval more natural and more appealing, but generally has little notion of how the two concepts are related.

2. An Example.

Let us review both procedures with reference to a numerical example.

For a certain type of shell, specifications state that the amount of powder should average 0.735 lb. In order to determine whether the average for the present stock meets the specification, twenty shells are taken at random and the weight of powder is determined. The sample average ($\overline{X}$) is 0.710 lb. The estimated standard deviation (s) is 0.0504 lb. The question is whether or not the average of present stock differs from the specification value. In order to do a two-sided test of significance at the $(1-\alpha)$ probability level, we compute a critical value, to be called for example, C. Let $C = \frac{t^* s}{\sqrt{n}}$

where t^* is the positive number exceeded by $100\left(\frac{\alpha}{2}\right)\%$ of the t-distribution with $n-1$ degrees of freedom.

In the example above with $\alpha = .05$, $t^* = 2.09$, $C = 0.0236$ lb. The test of significance says that if $|\overline{X} - 0.735| > C$, we decide that the average for present stock differs from the specified average. Since $|0.710 - 0.735| > 0.0236$, we decide that there *is* a difference.

We can also compute from the data a 95% confidence interval for the average of present stock. This confidence interval is $\overline{X} \pm C = 0.710 \pm 0.0236$ or 0.686 to 0.734 lb. The confidence interval can be used for a test of significance; since it *does not include* the standard value 0.735, we conclude that the average for the present stock *does differ* from the standard.

Comparisons of two materials (both means unknown and equal variances) may be made similarly. In computing a test of significance we compare the observed difference $|\overline{X}_A - \overline{X}_B|$ with a C′ (a computed critical quantity similar to C above). If $|\overline{X}_A - \overline{X}_B|$ is larger than C′ we declare that the means differ significantly at the chosen level. We also note that the interval $(\overline{X}_A - \overline{X}_B) \pm C'$ is a confidence interval for the difference between the true means $(\mu_A - \mu_B)$. If then this interval does not include zero, we conclude from the experiment that the two materials differ in mean value.

3. Do the Two Approaches Differ?

Here then are two ways to get the same answer to the

original question. We may present the result of a test of significance, or we may present a confidence interval. Are there any differences between the two? The significance test is a "go no-go" decision. We compute a critical value C, and we compare it with an observed difference. If the difference exceeds C, we announce a "difference"; if it does not, we announce no "difference." If we had no OC curve for the test, our decision would be a yes-no proposition with no shadowland of indifference. The test may say NO, but only the OC curve can qualify this by saying that this particular experiment had only a ghost of a chance of saying YES to this particular question. For example, see Fig. 2. If the true value $d = \left|\frac{\mu_1 - \mu_0}{\sigma}\right|$ is equal to 0.5, a sample of 10 is not likely to detect a difference, but a sample of 100 is almost certain to.

Using a rejection criterion alone is not the proper way to think of a significance test. One should always think of the associated OC curve as part and parcel of the test. Unfortunately this has not always been the case, and the significance test without its OC curve has distorted the thinking in some experimental problems. As a matter of fact many experimenters who use significance tests are using them as though there were no such thing as an OC curve. For this reason, it may be preferable for the experimenter to approach the problem of testing hypotheses by using confidence intervals.

4. Why Prefer the Confidence Interval?

A confidence interval procedure contains information similar to the appropriate OC curve, and at the same time is intuitively more appealing than the combination of a test of significance and its OC curve. If the standard value is contained in the confidence interval, one can announce "no difference." The *width* of the confidence interval gives a good idea of how firm is the Yes or No answer.[1]

Suppose that the standard value for some property is known to be 0.735, and that a $100(1-\alpha)\%$ confidence interval for the same property of a possibly different material is determined to be 0.600 to 0.800. It is true that the standard value is in the interval, and that we would say that there is no difference. All that we really know about the new product, however, is that its mean probably is between 0.6 and 0.8. If a much more extensive experiment gave a $100(1-\alpha)\%$ confidence interval for the new mean of $0.60 - 0.70$, our previous decision of no difference would be reversed.

On the other hand, if the computed confidence interval, for the same confidence coefficient, had been $.710 - .750$, our answer would still have been "no difference", but we would have said "No" more loudly and firmly. The confidence interval not only gives a Yes or No answer, but also, by its width, gives an indication of whether the answer should be whispered or shouted.

This is certainly true when the width of the interval, for a given confidence coefficient, is a function only of n and the appropriate dispersion parameter (e.g., known σ). When the width itself is a random variable (e.g., is a fixed multiple of s, the estimate of σ from the sample), one can occasionally be misled by unusually short or long intervals. But the *average width* of the entire *family* of intervals associated with a given confidence-interval procedure is a definite function of the appropriate dispersion parameter, so that *on the average* the random widths do give similar information. See [1] for a graphical illustration of confidence intervals computed from 100 random samples of $n = 4$ (actually random normal deviates).[2] Figure 14 in reference [6] shows a similar illustration of 100 intervals for $n = 4$, and in addition shows 40 intervals for $n = 100$, and 4 intervals for $n = 1000$. The fluctuation in size and position is of course very much reduced in the latter cases.

The significance test gives the same answer, and a study of the OC curve of the test indicates how firm is the answer. If the test is dependent on the value of σ, the OC curve has to be given in terms of the unknown σ, In such a situation, one has to make use of an upper bound for σ in order to interpret the OC curve, and again one may be misled by a poor choice of this upper bound. On the other hand the *width* of the confidence interval is part and parcel of the information provided by that method. No *a priori* estimates need be made of σ as would be necessary to interpret the OC curve. Furthermore, a great advantage of confidence intervals is that the width of the interval is in the same units as the parameter itself. The experimenter finds this information easy to grasp, and to compare with previous information he may have had.

5. What does the Confidence Interval Show?

The most striking illustration of information provided by confidence intervals is shown in the charts of confidence limits for a binomial parameter. In this case the limits depend only on n and the parameter itself, and one cannot be misled in an individual sample. Figure 1 shows the "central" 95% confidence limits for proportions. These "central" limits are the well-known Clopper-Pearson limits, such that each tail probability is not greater than .025. The central limits correspond to an equal-tail significance test at the $(1-\alpha)$ probability level, and to each of the two "central" limits there corresponds a single-tail significance test at the $(1-\alpha/2)$ probability level. In constructing a system of confidence limits there is no unique method of subdividing between the two tails. Limits which are not "central" may have other optimum properties—e.g., the recently-developed system of E. L. Crow [4] gives limits which are shorter than the "central" limits.

Suppose that a new item is being tested for comparison with a standard. In a sample of 10 we observed two

[1] There is a caution in this regard as explained a little further on.

[2] This picture is an excellent teaching aid in itself. Despite the fluctuation in size and position of the individual intervals, a proportion of the intervals remarkably close to the specified proportion do include the known population average. If σ were known rather than estimated from the individual sample, the intervals would fluctuate only in position, of course.

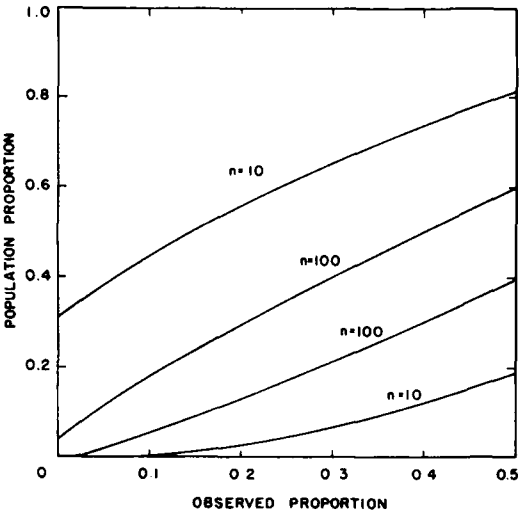


Fig. 1 95% Confidence Limits for Population Proportion

defectives and therefore estimate the proportion defective for the new item as 0.20. The central 95% confidence interval corresponding to an observed proportion of 0.20 ($n = 10$) is $0.02 - 0.56$. Assume that the known proportion defective for the standard (P_0) is 0.10. Our experiment with 10 gives a confidence interval which includes P_0, and therefore we announce "no difference" between the new item and the standard in this regard. Intuitively, however, we feel that the interval $0.02 - 0.56$ is so wide that our experiment was not very indicative. Suppose then we test 100 new items and observe 20 defectives. The observed proportion defective is again 0.20. The confidence interval now is $0.13 - 0.29$, and does *not* include $P_0 = 0.10$. This time we are forced to announce that the new item "is different" from the standard, and the narrower width of the confidence interval ($0.13 - 0.29$) gives us some confidence in doing so.

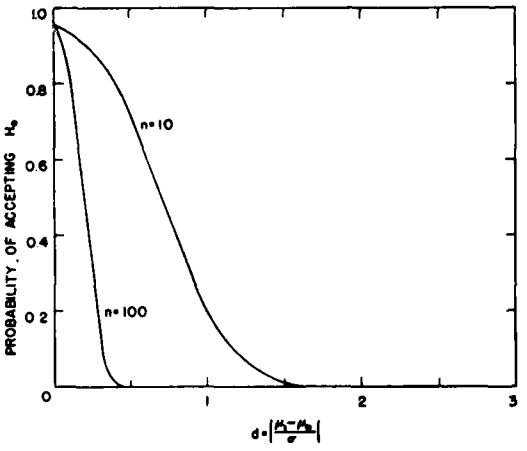


Fig. 2 Operating Characteristics of the Two-Sided t Test ($\alpha = 0.05$)

6. What does the Operating Characteristic Curve Show?

The foregoing has shown that it is possible to get some notion of the discriminatory power of the test from the size of confidence intervals. Is it also possible in reverse, to deduce from the OC curve what kind of confidence interval we would get for the new mean? Although we cannot deduce the exact width of the confidence interval, we can infer the order of magnitude. Suppose that we have measured 100 items, have performed a two-sided t-test (does the average μ_1 differ from μ_0?), and have obtained a significant result. Look at the curve for $n = 100$ in Figure 2, which plots the probability of accepting H_0 (the null hypothesis) against $d = \left| \frac{\mu_1 - \mu_0}{\sigma} \right|$. From the curve we see that, when *d* is larger than 0.4, the probability of accepting the null hypothesis is practically zero. Since our significance test *did* reject the null hypothesis, we may reasonably assume that our $d = \left| \frac{\mu_1 - \mu_0}{\sigma} \right|$ is larger than 0.4, and thus perhaps infer a bound for the true value of $|\mu_1 - \mu_0|$, in other words, some "confidence interval" for μ_1.

On the other hand, suppose that only 10 items were tested and a significant result was obtained. If we look at the curve for $n = 10$ in Fig. 2, we see that the value of *d* which is practically certain to be picked up on a significance test is now $d = 1.5$ or larger. A significant result from an experiment which tested only 10 items thus, as expected, corresponds to a wider confidence interval for μ_1 than that inferred from the test of 100 items. A rough comparison of the relative widths may be made. More quantitative comparisons could be made, but the purpose here is to show a broad general relationship.

7. Relation to the Problem of Determining Sample Size.

The problem of finding the required sample size to detect differences between means can also be approached in two ways. We can specify tolerable risks of making either kind of "wrong" decision (errors of the first and second kind)—thereby fixing two points on the OC curve of the required test. Matching these two points with computed curves for various *n* enables one to pick the proper sample size for the experiment.

Alternatively, we can specify the magnitude of difference between means which is of importance. We then compute the sample size required to give a confidence interval of fixed length equal to the specified difference.

8. Conclusion.

Presentation of results in terms of confidence intervals is often more meaningful than is the presentation of the usual tests of significance (if the test result is not considered in connection with its OC curve). Things are

rarely black or white, and decisions are rarely made on one-shot tests, but usually in conjunction with other information. Confidence intervals give a feeling of the uncertainty of experimental evidence, and (very important) give it in the same units, metric or otherwise, as the original observations.

The author is indebted to Dr. Churchill Eisenhart and Dr. Norman C. Severo who encouraged the writing of this paper.

REFERENCES

[1] *ASTM Manual on Quality Control of Materials*, (1951), p. 45. Available from the American Society for Testing Materials, 1916 Race Street, Philadelphia, Pennsylvania.

[2] Joseph Berkson, "Comments on Dr. Madow's 'Note on Tests of Departure from Normality' with Some Remarks Concerning Tests of Significance", *J. Amer. Statist. Assoc. 36*, No. 216, pp 539-541 (1941).

[3] Joseph Berkson, "Tests of Significance Considered as Evidence", *J. Amer. Statist. Assoc. 37*, No. 219, pp 325-335 (1942).

[4] Edwin L. Crow, "Confidence Intervals for a Proportion", *Biometrika, 43* pp 423-435 (1956).

[5] Maurice G. Kendall, *The Advanced Theory of Statistics, Vol. II*, pp 134-136, Hafner Publishing Co., New York (1951).

[6] Walter A. Shewhart, edited by W. Edwards Deming, *Statistical Method from the Viewpoint of Quality Control*, The Graduate School, Department of Agriculture, Washington, D. C. (1939).

[7] W. Allen Wallis and Harry V. Roberts, *Statistics, A New Approach*, pp 461-463, The Free Press, Glencoe, Ill. (1956).

Computations with approximate numbers[1]

The Mathematics Teacher | November, 1958

D. B. DELURY, *Ontario Research Foundation, Ontario, Canada.*

Contrary to present practices in the schools, there are no simple answers and no general rules for computing with approximate numbers.

THERE IS ROOM, I think, for the view that it is improper to speak at all of "approximate numbers." Admittedly this is jargon, an abbreviated phrase that we use to spare us the effort of saying over and over exactly what we mean, namely, "a number whose value approximates the value of this or that magnitude." There is no denying the usefulness of jargon. As long as it is used within the circle of experts or specialists for whom it is intended, there is probably little danger of confusion. On the other hand, there is no doubt that it can cause serious misunderstandings when it is used outside these circles. We have all seen the weird interpretations put on such words as infinity, curved space, the fourth dimension. It seems to me that "approximate number" is especially exposed to misinterpretation, because we have taken the word "approximate" from the place where it belongs, as a description of the *process* that produced the number, and attached it to the number itself.

I incline to the view that those of us who are concerned with giving accurate instruction should avoid this kind of jargon or, at least, delay its introduction until the need for it and the precise meaning to be attached to it have become amply clear.

When we come to speak about computations with such numbers and more particularly about the proprieties concerning the presentation of the results of these computations, it is, I believe, important to distinguish sharply between the two ways in which questions of approximation can arise:

1. through arithmetical operations only;
2. through measurement.

For the present, I shall speak of 1.

Since calculations for the most part are conducted with digital numbers, we are compelled to use approximations to irrational numbers and usually we find it convenient to use decimal approximations to rational numbers also. Here there ought to be general agreement on the procedures to be followed in truncating numbers, not because there is only one way or because there is necessarily one best way, but simply because, if only one rule is used, it need not be stated over and over.

If, for example, I wish to use a modest approximation to e, I look up its value in a book and find $e = 2.718281 \cdots$. Now I could truncate it simply by chopping it off at the desired place, e.g., $e = 2.71$. Then, I am required to take the view that e lies somewhere in the range 2.71–2.72. Let us make no mistake here; these numbers are exact and should be so treated. They simply mark the boundaries of a range within which *e certainly lies*. If I want to make calculations with the value of e so specified and to know within what range

[1] An address delivered at the 1958 Annual Easter Meeting of the Mathematics and Physics Section of the Ontario Education Association. It is being reprinted with the permission of the *Canadian School*.

the true answer certainly lies, I must make the calculation with each end of the range, thus obtaining the range within which the true answer must certainly lie. These calculations should be exact, to the extent that they can be, unless we wish to add arithmetical mistakes to the uncertainty caused by the truncation. Thus e^2 lies in the range $(2.71)^2$–$(2.72)^2$ = 7.3441–7.3984.

The notion of a range within which a number must certainly lie is the foundation on which every accurate procedure for computation must be based. Indeed, some writers use the term "range number" to give the idea a name. Range numbers need not arise only through the kind of truncation I used in the little illustration. I suppose there are all sorts of ways in which we could come by the information that "this magnitude certainly lies within the range so and so." If we were prepared to use range numbers always and perform the dual calculations they require, there would be no need here to go any further, beyond developing some simple rules to assist in performing quickly the fundamental arithmetical operations with range numbers. The fact is, however, that rarely are we willing to do the work that is required to make a calculation with range numbers. We prefer to sacrifice both accuracy and clarity in order to be able to calculate with single numbers. In the example we have been using, we could accomplish this by replacing the range number by the value at the middle of the range, 2.715, thereby rendering the maximum inaccuracy as small as possible. Then $e^2 = (2.715)^2 = 7.371225$. One might complain that this answer is wrong. I would prefer to say that it is not complete. To make it complete, we would have to write

$$e = 2.715 \pm .005,$$

whence

$$e^2 = (2.715)^2 \pm 2(.005)(2.715) + (.005)^2$$
$$= 7.371225 \pm .02715 + .000025.$$

Hence e^2 lies in the range

7.344100–7.398400.

This is, of course, just another way of using the notion of range. The result is, and must be, the same whichever way we choose to write our numbers. In particular, the zeros that came out in this calculation may be retained or dropped, at our pleasure. We may remark, too, that even though one of these calculations uses more digits than the other, their accuracies are the same. Numbers written in the form $— \pm —$ are sometimes called approximation-error numbers.

This second calculation is somewhat more tedious than the first, because the numbers we had to use require more nonzero digits. This stems from the way in which we performed the truncation. It gave us nice end points, 2.71 and 2.72, but a nasty mid-point, 2.715. If we had intended to use the second form of computation, we would have done better to use a form of truncation which requires as few digits as possible for the chosen accuracy, for the mid-point. We would, presumably, have said "e lies in the range 2.715–2.725," which has, it is true, nasty end-points but a nice mid-point, 2.72. We would then write e as the approximation-error number $2.72 \pm .005$. The largest possible inaccuracy is the same as in the first truncation.

It appears then, that if we propose to do our arithmetic with range numbers, we would use the first way of performing the truncation, and if we want to use the approximation-error form we would use the second.

Now in fact, we do not, as a rule, want to do our arithmetic in either of these ways. We want to calculate with the midpoints and forget about the inaccuracy. This point of view has dictated the universal practice of truncating in the second way, rather than the first. This process we call "rounding-off" and the resulting number we call a "significant number." A significant number, then, is one whose maxi-

mum inaccuracy is $\pm\frac{1}{2}$ in the last recorded digit. Significant numbers have the property that, for a given range of inaccuracy, they require fewer digits to specify exactly the mid-point than one would get by any other way of performing the truncation.

Since the convention about rounding off is, for all practical purposes, universal, it is generally understood that a truncated number is a significant number, and hence it is not necessary to specify the range of inaccuracy, since it is known to be $\pm\frac{1}{2}$ in the last recorded digit. This has some advantages. We know, for example, that the numbers listed in our tabulations of values of functions are significant numbers. On the other hand, the notion of significant numbers probably played no part in the computations that produced them. Significant numbers are not good numbers to calculate with, because the result of any computation with a significant number is not a significant number. For example, $e=2.72$ is a significant number, i.e., e certainly lies in the range $2.72\pm.005$. Hence, we can calculate that e^2 certainly lies within the range $(2.715)^2$–$(2.725)^2$ $=7.371225$–7.425625. There is no significant number equivalent to this range number. The best we can do is 7.4, which yields a range 7.35–7.45. Now it is true that this range certainly contains e^2, but it is far wider than it has to be in order to have this property.

This kind of mistake snowballs rapidly in extended calculations. For this reason, people who take their computations seriously do not use significant numbers, nor do they necessarily state the results as significant numbers. Hence, today we never know whether the numbers we see are significant numbers or not, except in books of tables. There is no way of telling, from the look of a number, whether it is a significant number or not. Things are, to put it mildly, somewhat confused. There is really no need for this situation, either. Largely it springs from the attempt to use significant numbers in dealing with measurements about which I shall speak later. If we stay for the moment within the field of arithmetic, which is the only place where the notion of significant number has any meaning, there is no need whatever for any confusion.

Let us agree that any number that has been reached solely through the operations of arithmetic, if it cannot conveniently be stated exactly, *ought* to be written as a significant number if possible; if not, there should be an explicit statement of its accuracy.

Presumably, before we embark on such a calculation, we know how many digits we wish to have in the result. The only question, then, that needs answering is "How shall we carry out the calculation in order to get the result we want?" It appears that we have been asking our question the wrong way round. We have asked "How should we calculate with approximate numbers?" when it would have been better to ask "How should we proceed to obtain an approximate number with the desired degree of accuracy?"

Having asked our question thus, it must be admitted that there is no simple answer and there are no general rules. We can, however, easily trace through the effects of arithmetical inaccuracy, caused by truncation of numbers, in a single application of each of the fundamental operations of arithmetic, and these effects can, and have been, formulated as rules.

Perhaps I should stop here for a moment to gather up the substance of what I have been saying.

1. I have been talking only about arithmetic.
2. Exact computations with approximate numbers are quite feasible, if somewhat distasteful, using range numbers. Obviously such calculations can lead to numbers with many digits. There is nothing improper in this. Indeed, the number of digits has nothing to do with the question of accuracy. However, these numerous digits can be a nuisance and we might well adopt the position

that, in view of the range, perhaps large, within which the answer lies, it is not important that we know exactly the boundaries of this range. We would, therefore, round off our numbers at some chosen number of *decimal places*. The choice would be determined by the use to which our answer is to be put, not by any considerations of inherent accuracy.

3. A significant number is, by definition, one that cannot differ from the one to which it approximates, by more than $\frac{1}{2}$ in the last recorded digit. Significant numbers are formed by the process of truncation called rounding-off. Significant numbers are, or if you prefer, easily lead to, range numbers, but range numbers can rarely be written as significant numbers. The convention that numbers which result from *arithmetical operations only* be written as significant numbers is useful and should be maintained quite generally. It may be remarked, though, that these circumstances do not arise as often as might be supposed. Most of our calculations are made with numbers that we get from measurements and to these the notion of significant number does not apply.
4. Significant numbers are not well suited to numerical calculations and are not so used by people who take their arithmetic seriously.

In connection with this last item, the following quotations are pertinent:

> The loss in the number of significant figures in products and quotients, for example, is due not so much to accumulation of errors as to the simplicity that has been gained at the expense of precision.

By *simplicity*, here, must be meant *ease of computation*. Actually, the use of significant numbers in computation makes for considerable complexity. Here, for example, is the theorem for products and quotients of significant numbers.

> The product or quotient of two numbers, each containing n significant figures (at least two of which are not zero) is a significant number of at least $(n-2)$ figures. If the leading digits of these numbers are both equal to or greater than 2, then the product or quotient has at least $n-1$ significant figures.[2]

This, I suggest, is not a step in the direction of simplicity.

> The selection of a suitable type of approximate number depends on the purpose of the computation. Operations with significant numbers are easier and simpler than the corresponding operations with range or approximation-error numbers. They are quite satisfactory when additions, subtractions or a single multiplication or division are involved. They are also satisfactory when we are not concerned with the loss of significant figures in each operation. In most computational work we cannot afford this luxury.[3]

We can take it, then, that whenever we can avoid the use of range or approximation-error numbers, we will do so, but we prefer not to use significant numbers. What, then, do we do? One more quotation from the same book:

> For the basic linear problems the method of the next paragraph is to be preferred.
>
> An alternative method is the use of *incomplete* numbers. An incomplete number is an approximation-error number in which the error term is omitted. These numbers look very much like significant numbers, but, unlike significant numbers, the results may be recorded to any desired number of places. This method makes for ease with a machine, since all numbers to be placed on the machine may be rounded off to the same number of places. It must be remembered that any recorded number is not necessarily a significant number in the technical sense, that is, we do not know what the bound for the error may be.[4]

This is surely a curious statement, with its invention of the term *incomplete* number. It means simply this, that every number that enters into the calculation is treated as if it were exact, to a chosen number of decimal places. The number of decimal places is chosen to be adequate, or usually much more than adequate, to yield results of the required accuracy. This is the way computers usually do their arithmetic.

[2] Dwyer, *Linear Computations* (1951), p. 15.
[3] *Ibid.*, p. 33.
[4] *Ibid.*, p. 34.

Let us pass on now to the second way in which we come to deal with numbers that are approximations to others—those which arise from measurement. Let us restrict ourselves to physical measurements—the so-called measurements of psychology and such fields pose somewhat special problems.

To start the discussion on this topic, I offer you a quotation from N. R. Campbell: "Probably more nonsense is talked about measurement than about any other part of physics."

I acquired this quotation second-hand and I cannot say that he had in mind the kind of thing I must talk about today. If he didn't, I shall make the same statement in this context.

In the first place, measurement does not produce numbers. The result of a measurement should properly be stated in the form of a range, within which some point, line, or whatever is observed to lie. Presumably this process could be carried out in such a way that this range could be expressed as a number, plus or minus $\frac{1}{2}$ in the last recorded digit. This is not always done, by any means, but let us say that it is. Then, it seems to be an easy step to say that these recorded numbers, obtained from measurement, are *significant* numbers, that is, the numerical value of the magnitude, which our process of measurement is supposed to estimate, *certainly lies* within the range defined by $\pm\frac{1}{2}$ in the last recorded digit.

It is a pity that this is not true. If it were, the world of science would be much simpler than it is. The fact is, of course, as we all know, measurements don't behave this way at all. We all know, for example, that the real error of any measurement is not composed wholly of the error committed in making a final scale reading. Not uncommonly, two attempts at measuring the same thing produce two ranges which do not overlap at all and we cannot, surely, be certain that the true value lies in both of them.

This implies that the real error is likely to be far greater than that implied by the coarseness of the scale with which the final reading is made. Often it is so great that this final contribution to the error may be ignored. On the other hand, repeated determinations, made with even a coarse scale, can lead to an estimate which is far more precise than the coarseness of the scale would lead one to expect. What I am saying, then, is simply this: no number, significant, range, or any other, can, by itself, say anything about the precision or accuracy of the measuring process used to obtain it.

Now, this has been known for a long time. Over 100 years ago, Gauss and Laplace worked over this ground and produced the theory of errors, based on the notion of the frequency distribution, which is simply an idealization of the experience of people who make measurements on how errors of measurement behave. They did the job pretty well, too, and this theory has come down to us virtually unchanged, except that a comparatively new discipline, the design of experiments, has given it considerably greater depth and scope.

In any event, the theory of errors provides us with the only usable tool we have for dealing with errors of measurement. This tells us, among other things, that the treatment of errors cannot be undertaken as an exercise in arithmetic. The basic requirement is that the program of measurement be so arranged as to permit the estimation of a *standard deviation*, in terms of which the precision of the measuring process can be stated. The statement that the length of something or other is 11.3 inches, to the nearest tenth-inch, is either trivial or not true. The statement that the length of something or other is 11.3 inches, with a standard deviation of .2 inch, *is* a meaningful statement. It implies, among other things, that the "true" length is almost certainly somewhere between 10.7 and 11.9 inches, leaving aside the possibility of bias or systematic error, which is irrelevant to our

topic. Note, however, that this is *not* a range number, because we cannot know that the true value is *certainly* in this range. To say the same things in another way, there is no place, in connection with errors of measurement, for any talk about *possible* errors, meaning the maximum error that could possibly be encountered. The notion of possible error, or, I would prefer to say, possible *inaccuracy*, has meaning only in arithmetic.

I do not know how the notion has crept in that the theory of errors can be replaced by an exercise in arithmetic, coupled to a convention about the form in which the answer should be written, but I favor the view that the physicist is the culprit, not because I know of anything in the literature that points a finger at him, but rather because physicists have generally been loath to carry out their programs of observations in such a way that their real errors can be estimated and the theory of errors brought properly into play. In these circumstances, then, they have tried to assign what they call a "limit of error," that is to say, a maximum possible error. Let me quote from a book recently published, written by a physicist, on the Theory of Error:

> Many observers estimate the limit of error, the maximum amount by which the quantity may be supposed to be in error. Other observers believe such a procedure too conservative, since large errors are relatively improbable compared with small ones. Therefore, instead of using the full estimated value of the limit of error, these observers reduce it, perhaps by one-third. Since these are matters of opinion, no firm rules can be given and each experimenter must use his own judgment.[5]

This says to us, it seems, "Pick the error you like best."

Now, the theory of errors is not a matter of opinion and there is no doubt at all about how an experimenter should carry out his work in order to make proper use of it. This kind of humbug is an attempt to gain an air of respectability by sneaking under the mantle of the theory of errors without doing the work that it demands. With respect to practices of the kind revealed in this quotation, a remark made by Bertrand Russell in his *Introduction to Mathematical Philosophy* seems to me to be wholly pertinent.

> The method of postulating what we want has many advantages; they are the same as the advantages of theft over honest toil.[6]

Presumably this is not the place for an exhortation about the importance of the theory of errors. My whole concern here is to bring out one fact, that errors of measurement have nothing whatever in common with so-called approximate numbers. They are conceptually wholly different. One is concerned entirely with a question in arithmetic, specifically, with the magnitude of the mistakes (rather than errors) that can enter into an arithmetical calculation through truncation of numbers. The other has to do with the physical processes of making measurements and is based on empirical evidence about the way measurements behave. In the one, it is wholly proper to speak of the largest possible *mistake* introduced into an arithmetical calculation by truncation of numbers; in the other, it is quite improper to speak of a maximum possible *error* in a measurement. It is important, if we want to think accurately about these matters, that we keep these two things sharply separated.

Where, then, does all this leave us? Specifically, what does this mean for us, who are concerned with providing instruction that is accurate and comprehensible at the level at which we must give it?

I would not undertake to give any comprehensive answer to this question. Indeed, I believe that for me or anyone else to do so, without having tried and experimented a bit to build up some experience on what can be done successfully and what can't, would be sheer folly. On the

[5] Yardley Beers, *Introduction to the Theory of Error* (Cambridge, Mass.: Addison-Wesley Publishing Company, Inc., 1953).

[6] Bertrand Russell, *Introduction to Mathematical Philosophy* (London: George Allen and Unwin, Ltd., 1919).

other hand, there are certain principles which must be rigorously observed in *any* attempts in this direction. These have, for most part, already emerged, but I shall list them again and offer some opinions on the implications of these principles.

1. Errors of measurement should be sharply distinguished from mistakes in arithmetic of the sort that lead us to speak of significant numbers. I believe that this separation should be so stoutly maintained that no discussion of measurement would be permitted in mathematical subjects. In these, and especially in trigonometry, we have a good opportunity to see the way in which arithmetical calculations are properly carried out, with due attention paid to orderly procedures, mistakes introduced by truncation, and checks against blunders. After all, there is only one reason why anybody performs an arithmetical calculation. It is to get the right answer. As far as the question of approximate numbers is concerned, there is room here, I should think, for studying how truncation mistakes are propagated through simple calculations. I would make no room at all for horrible examples in which numbers with grossly different accuracies have to be combined. If the arithmetic is handled competently, this should not be allowed to happen. The emphasis should be on how one should carry out his calculation to arrive at a result of the required accuracy. It is, in my opinion, quite proper to put a problem in the following form: Two sides of a triangle are 9 metres and 12.3 metres long. The angle between them is 27°39′. Calculate approximately the area of the triangle and give a range within which the area of the triangle must lie.

The numbers given in this question are, by implication, exact.

It is true that there are many answers to this question, all of them correct. This is as it should be. If we want to single out a particular one of these, we can require the answer to so many significant figures or, what comes to the same thing, we can ask for the area, in square metres, correct to so many places of decimals.

One might object that the question so stated is artificial. If, by this, is meant that if one had to measure these quantities, he could not ask such a question, I agree. On the other hand, it is a real *mathematical* question and I remind you that I am talking about a course in mathematics. Furthermore, the question is no less artificial, from the point of view of measurement, if it is asserted that the sides are measured to the nearest centimetre and the angle to the nearest minute, or if the same notion is conveyed by adopting the convention of significant numbers and writing 9.0 instead of 9, and so on. It is, I repeat, no less remote from reality and, in addition, it carries the implication that we can cope with errors of measurement in this manner, which is monstrous. If we allow ourselves to go this far, we might as well go all the way and write our lengths as 9.0 and 12, because now they have the *same number* of *significant figures*. At this point, we are as far from the realities of measurement as we can get.

If we want to ask such a question in the only proper way it can be asked, when the dimensions of the triangle are measured, it would have to read somewhat as follows: The sides of a triangle are 9 and 12.3 metres, with a standard error of .12 metres; the included angle is 27°39′, with a standard error of 5′. Calculate the area of the triangle and its standard error.

This is no mean problem!

2. The basic notion in any discussion of mistakes in arithmetic, caused by truncation, is the range number. It is so simple a concept that any child can see all that is involved and it seems to me the natural place to begin. The significant number then emerges as a special kind of range number and its merits and weaknesses are immediately obvious. The dogma of the significant number has been with us for a long time and I fear it will plague us for some time to come, so we can hardly avoid discussing it in our teaching. However, let

us teach it as it is, not with a halo of false implications surrounding it, and we should make it clear that we do not encounter genuine significant numbers very often.

All in all, it seems to me that what needs doing and a general outline of how to do it are fairly clear, as far as purely arithmetical questions are concerned. It may be granted that the going gets rough when we come to elaborate calculations, such as the solution of large systems of equations, but such questions should not arise at an elementary level.

3. Things are less clear when we turn to errors of measurement, but it seems natural to me to suppose that the place to talk about measurements is where measurements are made, at least where they are made seriously, to learn something about a physical or chemical system. I do not regard as acceptable here the sort of exercise that has gained favor in some places, for example, sending a number of pupils with yardsticks to measure the length of a room, taking their measurements, which may read 24′, 23′9″, 23′7¼″, and so on, then using these numbers to show how to reduce them all to the level of the worst, average them, and come out with an estimate of the length of the room. I can see nothing but harm in such exercises. Not only do they invoke the confusion of approximate numbers with errors of measurement, but they cast a false light on the process of measurement itself. In all honesty, anyone who treats numbers so obtained as anything but garbage makes a mockery of the whole idea of measurement. The making of measurements is too serious a business to be treated so casually. It is, among other things, a complex physiological and psychological process and is without meaning until stability and control have been demonstrated.

Really, I see nothing to be gained by talking about errors of measurement, except to people who make measurements, that is, in courses in physics, chemistry, and perhaps biology. Maybe laboratory work can be planned to give some indication of the way measurements behave. Measurements must be repeated and, indeed, if one is concerned to know the whole of his error, whole experiments must be repeated.

I do not suggest that it is desirable or feasible to introduce any discussion of the theory of errors at the high school level. It is a topic that apparently demands considerable maturity. Even in universities, no serious attempts are made to provide the rudiments of the theory of errors to all the people who need them. However, it should be quite feasible to discuss the notion of bias (the systematic error of the physicist) and that of true errors, which tend to compensate. In connection with these, it is vitally important to give a careful and thoughtful discussion of the notion of average, what it can accomplish and what it can't, when one has the right to use an average and when he hasn't. I believe that, today, the average is among the most overworked and most abused of all our concepts.

4. Above all, we must scotch the notion that the precision of an average or of a single measurement can be judged from the way in which it is written. Apparently the opinion is current that some impropriety attaches to running an average out to "more places than are warranted," indeed, that there is some suggestion of deception in that more precision is claimed than can be justified. Now, we may grant that running averages of measurements out to many places of decimals is silly, but the real impropriety lies not in this, but in failing to provide a standard error to go with this average.

Perhaps an example will gather up some of the notions I have been putting forward. The example comes unchanged from the A.S.T.M. *Manual on Presentation of Data*. (Incidentally, in this manual, which is wholly concerned with the presentation and interpretation of measurements, there is no mention of significant numbers or significant digits.)

The ten numbers presented in the table are the measured breaking-strengths, in pounds, of 10 samples of copper wire, taken to estimate the mean breaking-strength of wire from one production lot.

Specimens	Breaking-Strength $= X$	X^2
1	578	334,084
2	572	327,184
3	570	
4	568	
5	572	
6	570	
7	570	
8	572	
9	596	
10	584	341,056
	$5,752 = \Sigma X$	$3,309,232 = \Sigma X^2$

Average $= \Sigma X/10 = 575.2 = X$

$\Sigma X^2/10 = 330,923.2$

$(\bar{X})^2 = 330,855.04$

Subtract 68.16

Extract square root 8.26 = standard deviation.

Remarks

1. The numbers X may represent readings rounded off to the nearest unit, i.e., $\pm \frac{1}{2}$ in the last digit, but on the other hand they may not and without knowing the details of the measuring process it would be improper to assume that they do. (The fact that they all are even makes one wonder!) In any event, this is irrelevant. We will do the same things with these numbers, no matter how the readings were made.

2. The object in making these measurements is to estimate the mean breaking-strength of this lot of wire. The average of the observations is calculated as an estimator of this number. It is here calculated to one place more than those given in the data. This is in accord with A.S.T.M. recommended practice. There is positively no implication that the true mean lies between 575.15 and 575.25. A range within which it is *likely* to lie is calculable from the standard deviation.

3. In the calculation of the standard deviation (and indeed in the calculation of the mean also) the observed numbers are treated as exact. To do anything else would bring us out with no meaningful numbers whatever.

4. The standard deviation is given to two places more than the data. This also is recommended A.S.T.M. practice.

Let us look at these A.S.T.M. recommendations as they apply to this particular example. A small statistical calculation indicates that there is near-certainty that the interval 575.2 ± 8.94 straddles the true mean. In view of the width of this interval, it is likely that these decimals serve no useful purpose and 575 ± 9 would meet all the needs we have. Thus it appears that the A.S.T.M. rule has given us more places than we have any use for. This has happened because most of the error has come from sources other than the final scale reading. On the other hand, if most of our error *had* been so caused, as it might be if we were measuring the value of some physical constant, these decimals might be well worth having. We see, then, I think, the meaning of these rules. They are simple conventions which have been adopted to keep people from doing outrageously silly things. They have not been derived from any fundamental considerations. Furthermore, there could be circumstances, I think, in which I would choose not to follow them.

These considerations have a bearing on the advice we should give to students who make measurements in the laboratory and then make calculations with them. I assume that the laboratory work will not be carried out in such a way that the theory of errors can be applied and that the theory of errors will not have been taught anyway. In these circumstances, the error in the measurements and in the final calculated estimates cannot be known. Two questions require answers. How shall they carry out their arithmetic and how shall they present the results of their calculations?

The answers are easily given. In their arithmetic, all numbers are to be treated as exact, with the proviso that if the num-

ber of decimal places becomes unduly large, some of them may be eliminated by rounding off in the course of the calculation. The final answer should be rounded off to a reasonable number of decimals.

I am sure we would, all of us, like to have something more definite than this, but the fact is that there are no grounds for definiteness. And, after all, what difference does it make if one person runs his calculation out to, say, two more decimal places than another? As long as they have not stopped too soon, the basis for a preference between them is largely aesthetic. Of course, before we can adopt this indulgent view, we must get rid of the notion, to which we never had any right, that we are dealing with significant numbers. It is this notion that has led to the view that too many decimals in the answer imply deception. As far as deception is concerned, the shoe is entirely on the other foot. The contention that the result of a calculation, with numbers obtained from measurement, is to be interpreted as a significant number is practically certain to deceive. Let us look back at the example and, for purposes of illustration, treat the numbers as significant numbers. Then we certainly have the right to express the average of them as a significant number, to the same number of digits, i.e., 575. Then, since 575 is a significant number, the "true" value certainly lies in the range 574.5–575.5. This statement is simply not true.

To sum up, then, let us keep the significant number where it belongs, as a convenient convention for writing answers in pure arithmetic. It has no other use.

Reprinted from *The Mathematics Teacher*
Volume LI, No. 7, November, 1958

NATIONAL BUREAU OF STANDARDS
Statistical Engineering Laboratory
Washington, D. C. 20234

October 1967

SELECTED REFERENCES

David Hogben

The list of books in this section is provided for persons active in measurement science who are concerned with the analysis and statistical treatment of measurement data. The object is to provide the scientist unfamiliar with statistical concepts a short list of books that he may turn to for help in understanding techniques which will be of use to him. The list has been kept short intentionally, and many good books have been excluded. No attempt has been made to include books intended for research workers in other fields and no attempt has been made to include advanced texts primarily of interest to statisticians. Comments, usually in the form of quotes from the author's preface, are given to guide the reader in making his own selection.

A. General (and easy reading)

1. WILSON, E. B., An Introduction to Scientific Research, McGraw-Hill, 1952 (375 pp.) Paperback, $2.75.

   This book is unique. It embraces the fundamental principles and methods of scientific research without the loss of important details. Data collection and analysis are treated in chapters 7 thru 9.

2. YOUDEN, W. J., Experimentation and Measurement, National Science Teachers Association, Washington, D. C. (127 pp.) Paperback, $0.50.

   Primarily written for young scientists, but the concepts introduced and techniques demonstrated are clearly applicable in general.

3. MORONEY, M. J., Facts from Figures, Pelican Books, 1951 (472 pp.)

Paperback.

An amusing, common sense approach to basic statistics, written for the non-mathematician. "The book ranges from purely descriptive statistics, through probability theory, the game of Crown and Anchor, the design of sampling schemes, production quality control, correlation and ranking methods, to the analysis of variance and covariance."

4. WILKS, S. S., Elementary Statistical Analysis, Princeton University Press, 1951. (284 pp.)

A nice blend of elementary methods and theory. "An effort has been made throughout the book to emphasize the role played in statistical analysis by a sample of measurements and a population from which the sample is supposed to have arisen. ... Considerable attention is given to the application of sampling principles to the simpler problems of statistical inference such as determining confidence limits of population means and difference of means, making elementary significance tests, testing for randomness, etc."

B. Statistical Methods

1. DAVIES, OWEN L., Statistical Methods in Research and Production, Hafner Publishing Co., New York, Third edition, 1957 (396 pp.)

"The object of this handbook is to bring together under one cover those methods of statistical analysis which are most likely to be of use in the Chemical Industry." Many useful topics are treated in detail and with worked examples.

2. YOUDEN, W. J., Statistical Methods for Chemists, John Wiley and Sons, 1951 (126 pp.)

This book provides a refreshingly elementary approach to many of the most basic problems in applied statistics. It is written for the experimenter in his language.

3. BROWNLEE, K. A., Statistical Theory and Methodology in Science and Engineering, Second Edition, John Wiley and Sons, New York, 1965 (590 pp.).

A middle ground between theory and applications.

4. DIXON, W. J., and MASSEY, F. J., Introduction to Statistical Analysis, McGraw-Hill Co., 1957, Second Edition (488 pp.).

An elementary but useful book.

5. HALD, A., Statistical Theory with Engineering Applications, John Wiley and Sons, Inc., 1952 (783 pp.).

"It is the aim of this book to provide a fairly elementary mathematical treatment of statistical methods of importance to the engineer in his daily work."

6. MANDEL, JOHN, Statistical Analysis of Experimental Data, John Wiley and Sons, Inc., 1964. $12.00 (410 pp.)

"The aim of this book is to offer to experimental scientists an appreciation of the statistical approach to data analysis." Most of the examples are "based on genuine data obtained in the study of real laboratory problems." Chapter 6, "The Precision and Accuracy of Measurements," is concerned with measures of experimental error. Some other topics particularly well presented are weighted averages, and the importance of a careful examination of residuals. Chapter 13, "The Systematic Evaluation of Measuring Processes," and the final chapter, "The Comparison of Methods of Measurement," are particularly interesting.

7. ACTON, FORMAN S., Analysis of Straight-Line Data, John Wiley and Sons, 1959 (267 pp.). Dover has a paperback edition.

This book is of limited scope, but it treats some problems which occur frequently with measurement data in considerable depth.

8. DRAPER, N. H. and SMITH, H., Applied Regression Analysis, John Wiley and Sons, 1966 (407 pp.).

"This book provides a standard, basic course in multiple linear regression, but it also includes material that either has not previously appeared in a textbook or, if it has appeared, is not generally available. For example, Chapter 3 discusses the examination of residuals; Chapter 6 examines the methods employed as selection procedures in various types of regression programs; Chapter 8 discusses the planning of large regression studies; and Chapter 10 provides a basic introduction to the theory of nonlinear estimation."

C. Manuals and Handbooks

1. ASTM Manual on Quality Control of Materials, ASTM Comm. E-11. Special technical publication 15-C, 1951 (127 pp.).

Part I - Presentation of data
Part II - Presenting ± limits of uncertainty of an observed average
Part III - Control chart method of analysis and presentation of data

(ASTM Address: American Society for Testing Materials, 1916 Race Street, Philadelphia, Pa. 19103)

2. NATRELLA, M. G., Experimental Statistics, NBS Handbook 91, U. S. Government Printing Office, Washington 25, D. C., 1963 (504 pp.) $4.25.

"The Handbook is intended for the user with an engineering background who, although he has an occasional need for statistical techniques, does not have the time or inclination to become an expert on statistical theory and methodology."

"Step-by-step instructions are given for attaining a stated goal, and the conditions under which a particular procedure is strictly valid are stated explicitly. An attempt is made to indicate the extent to which results obtained by a given procedure are valid to a good approximation when these conditions are not fully met. Alternative procedures are given for handling cases where the more standard procedures cannot be trusted to yield reliable results."

3. DUNCAN, A. J., Quality Control and Industrial Statistics, Third Edition, Richard D. Irwin, Inc., Homewood, Ill., 1965 (992 pp.).

This book presents "... the basic principles and procedures of statistical quality control. It is ... a discourse on the assumptions and principles of theory that underlie modern quality control practice."

Part I - Fundamentals
Part II - Lot Acceptance Sampling Plans
Part III - Rectifying Inspection
Part IV - Control Charts
Part V - Some Statistics Useful in Industrial Research

4. OWEN, D. B., Handbook of Statistical Tables, Addison-Wesley, 1962. (580 pp.)

A comprehensive and reliable collection of tables by an excellent table maker.

D. Design of Experiments

1. COX, D. R., Planning of Experiments, John Wiley and Sons, 1958 (308 pp.).

"This book is an account of the ideas underlying modern work on the statistical aspects of experimental design. I have tried, so far as is possible, to avoid statistical and mathematical technicalities, and to concentrate on a treatment that will be intuitively acceptable to the experimental worker, for whom the book is primarily intended."

2. YOUDEN, W. J., Statistical Design, American Chemical Society, Wash., D. C. (72 pp.).

Reprinted from Industrial and Engineering Chemistry, a collection of bimonthly articles (71) from 1954 to 1959.

3. HICKS, C. R., Fundamental Concepts in the Design of Experiments, Holt, Rinehart and Winston, 1964 (293 pp.).

A well written elementary book which presents "... a logical sequence of designs that fit into a consistent outline; for every type of experiment, the distinction among the experiment, the

design, and the analysis are emphasized."

4. DAVIES, OWEN L., Editor, The Design and Analysis of Industrial Experiments, Hafner Publishing Co., New York, 1954 (636 pp.).

"This handbook ..... is a sequel to Statistical Methods in Research and Production. It deals with ... the arrangement of the individual items composing a complex experiment designed for a given purpose, and the statistical analysis of the results. ... This book has been written principally for the research worker in the chemical industry with a limited knowledge of mathematical statistics, but it is hoped that it will appeal to other readers."

E. Textbooks on Probability and Statistics

1. ANDERSON, R. L., and BANCROFT, T. A., Statistical Theory in Research, McGraw-Hill Book Co., Inc., New York, 1952 (399 pp.).

"Many research workers have expressed a need for a convenient reference book on statistical theory pointed to research problems, which could be used in conjunction with their books on general statistical methods, experimental design, and survey sampling. The authors have tried to write a book which would serve this purpose as well as that of a textbook in statistical theory."

2. MOOD, A. M. and GRAYBILL, F. A., Introduction to the Theory of Statistics, McGraw-Hill Book Co., Inc., 1963 (443 pp.).

A standard advanced undergraduate/first-year graduate level text on the theory (rather than mathematics) of statistics having a one year of calculus prerequisite.

"While this text is primarily concerned with the theory of statistics, full cognizance has been taken of those students who fear that a moment may be wasted in mathematical frivolity. All new subjects are supplied with a little scenery from practical affairs, and, more important, a serious effort has been made in the problems to illustrate the variety of ways in which the theory may be applied."

7. Abstracts of Recent Publications

UNITED STATES DEPARTMENT OF COMMERCE • John T. Connor, Secretary
NATIONAL BUREAU OF STANDARDS • A. V. Astin, Director

TECHNICAL NOTE 288

ISSUED MAY 6, 1966

(Reprinted January 1968, with minor corrections)

Measurement Philosophy of the Pilot Program for Mass Calibration

P. E. PONTIUS

Metrology Division
Institute for Basic Standards
National Bureau of Standards
Washington, D.C., 20234

Contents

The Measurement Philosophy for the Pilot Program for Mass Calibration

Paul E. Pontius

The Pilot Program for mass measurement is the result of a consideration in which the values produced are thought of as the products of a mass measurement process. The collective performance of elements of the mass measurement process results in establishing the process precision which, under certain conditions, can be described quantitatively by pertinent performance parameters. The uncertainty attached to the product of the process, the measured value, is computed from these parameters and reflects the total performance of the process rather than the immediate measurement which might have produced the value. Interpretations of uncertainty and surveillance tests are discussed. The Pilot Program in mass measurement, whereby suitable process performance parameters can be established for precise mass measurement processes in other facilities, is discussed.

Key words: Mass measurement process, process performance parameters, and uncertainty.

1. Introduction

In order to utilize the capabilities of a particular mass measurement process, it is necessary to have at least one mass standard of known value to establish the measurement unit and, equally important, to know quantitatively how well the process performs. The process produces mass values for a wide variety of objects and, in most instances, the objects and values pass on to others to serve many purposes. The uncertainty associated with values produced by the process establishes the suitability of these values for the intended usage, the amount of measurement effort necessary to meet the requirements with confidence, and the basis for agreement when the same measurement must be made with two different measurement processes. If the uncertainty is to be realistic, it must be formulated from process performance parameters which are established by all the data generated by the process to date. In addition, it must adequately reflect both the random variabilities and systematic errors associated with the process.

The activities of the Mass and Volume Section and the Statistical Engineering Laboratory have been directed, for the past several years, toward an objective evaluation of the mass measurement process and toward the establishment of suitable parameters of performance which can be used to compute realistic estimates of process uncertainty to be associated with the mass values produced. The success of these efforts provide the basis for the formulation of a different method for disseminating the mass measurement unit and for maintaining the standards of mass which are directly involved in measurement processes throughout the country. The resulting program, currently designated the Mass Measurement

Pilot Program incorporates, in each participating facility, the calibration procedures currently in use at the National Bureau of Standards which provide both a means to recognize and to utilize the maximum capabilities of the mass measurement processes.

The Mass Measurement Pilot Program, at the present stage of development, requires the participating facility to either have, or have access to, a pair of kilogram mass standards and suitable sensitivity weights which have been recently calibrated by NBS. The calibration of duplicates, subdivisions, and multiples of the kilogram are accomplished by using the equipment of the facility to make the observations in accordance with the prescribed procedures. The raw data is transmitted to NBS via teletype or other convenient means of communication. The data will be processed using an appropriate computer program. The monitoring function incorporated in the analysis will test the values obtained for the performance parameters against the appropriate parameters which represent the performance of the facility which produced the data. The mass values and appropriate uncertainties for the weights being calibrated are returned to the facility via teletype in a format suitable for inclusion in a report of calibration. The analysis sheets, which include, in addition to the statistical eavluation, a listing of supplementary information such as the equipment used, the operator, the weighing designs used and also a copy of all of the raw data listed essentially in the order it was taken, are forwarded by mail for evaluation and use as substantiating documentation. At the present time, the program is in limited operation at three facilities over a restricted range of nominal mass values. The success of the operation, to date, has been most gratifying.

The programs for data analysis, incorporated in the Pilot Program, strive to provide a service matched to the unique requirements of the total mass measurement process and to extract from the resulting data all possible information concerning the process performance. The procedures are designed to calibrate most ordered sets of mass standards, with few if any, extra observations over those required by other calibration procedures, and in addition, one obtains the statistical information necessary to assess the performance of the particular process that was used. The analysis of the data provides parameters relative to both short and long term process variability and it is possible to compute in advance, and verify the appropriateness of the uncertainty to be associated with each mass value determined. Facilities that can demonstrate a continuous "in control" operation through the use of the Pilot Program are, in essence, extensions of the NBS facilities and, as such, require only minimal calibration support.

UNITED STATES DEPARTMENT OF COMMERCE
Alexander B. Trowbridge, Secretary
NATIONAL BUREAU OF STANDARDS • A. V. Astin, Director

NBS TECHNICAL NOTE 430

ISSUED OCTOBER 9, 1967

Designs for Surveillance of the Volt Maintained by a Small Group of Saturated Standard Cells

W. G. Eicke
Electricity Division
Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234

J. M. Cameron
Applied Mathematics Division
Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234

Designs for Surveillance of the Volt
Maintained by a Small Group of Saturated Standard Cells

W. G. Eicke
Electricity Division
Electrochemistry Section

and

J. M. Cameron
Applied Mathematics Division
Statistical Engineering Laboratory

ABSTRACT*

When a local standard such as that for electromotive force is maintained by a group of standards, procedures must be established to provide evidence that the group has maintained its original value. One also needs methods for the transfer of the value to test items that provide efficient use of measurement effort while monitoring the measurement process and providing information for updating the values of process parameters. Solutions to the more general problem of transferring the value from laboratory to laboratory and of maintaining agreement among laboratories depend on the existence of control within the laboratories.

This technical note describes a procedure for maintaining surveillance over a small group of saturated standard cells. The measurement process is briefly discussed and the principle of left-right balance as a means of eliminating certain systematic errors is developed. Specific designs and their analysis for intercomparing 3, 4, 5 and 6 cells in a single temperature control environment are given. Procedures for setting up control charts on the appropriate parameters are given, and a technique is described for detecting certain types of systematic errors.

Key words: Control charts, experiment design, saturated standard cells, standard cells calibration, statistics, voltage standard.

---

* Revised September 16, 1968

TABLE OF CONTENTS

UNITED STATES DEPARTMENT OF COMMERCE • John T. Connor, Secretary
NATIONAL BUREAU OF STANDARDS • A. V. Astin, Director

TECHNICAL NOTE 277

ISSUED July 25, 1966

Analytical Mass Spectrometry Section:

Instrumentation and Procedures for Isotopic Analysis

Edited by William R. Shields

Analytical Mass Spectrometry Section
Analytical Chemistry Division
Institute for Materials Research

TABLE OF CONTENTS

ANALYTICAL MASS SPECTROMETRY SECTION:

INSTRUMENTATION AND PROCEDURES FOR ISOTOPIC ANALYSIS

Edited by William R. Shields

ABSTRACT

This report describes the general instrumentation of the Analytical Mass Spectrometry Section and the specific analytical techniques which have been devised for the measurement of isotopic ratios of Ag, Br, Cl, Cr, Cs, Cu, Mg, Pu, and U. Interim procedures for B, Li, Rb, and Sr are also given.

In the appendix some general statistical principles used in the design and analysis are briefly discussed; an example is given in detail illustrating the various steps involved leading from original data to the reported uncertainties for the isotopic ratio of bromine.

Key words: Mass spectrometry, instrumentation, procedures, isotopic analysis.

Statistical Techniques for Collaborative Tests

W. J. Youden

CONTENTS

STATISTICAL TECHNIQUES FOR COLLABORATIVE TESTS

FOREWORD

This manual presents statistical techniques that may be used in collaborative testing of analytical methods. It is an introductory guide issued for use by the AOAC's Associate Referees, whose statistical backgrounds cover a wide range. Every effort has been made to keep the presentation simple and flexible, and to hold the statistics to a minimum. Special attention is focused on planning the collaborative test and presenting and interpreting the analytical results.

Obviously the laboratory proposing a new method should study the method carefully before a collaborative trial is undertaken. Section VI D of this handbook gives an efficient program for a within-laboratory examination of a method.

An important innovation is that materials be chosen in pairs (Section V). The two members of a pair should be similar in nature and amount present. Such a pair becomes a "unit block" in the collaborative tests. The concept of the unit block leads to an easy statistical analysis and graphical examination of the data. The data from several unit blocks are readily presented in summary form.

The concept of the unit block allows a way to utilize the services of smaller laboratories not having the resources to participate in some of the more comprehensive collaborative studies by having them participate on only the unit blocks of materials in the range of interest to them. It is not necessary that each unit block have the same number of collaborators. As a guiding rule, as many collaborators as possible should be obtained—up to 30 collaborators per unit block. At the minimum level it is suggested as a guide (not as a standard) that the number of collaborators be maintained at not less than six, if possible. We must recognize that situations sometimes exist which severely limit the number of collaborators that can be obtained. This rule is certainly not intended to cut off such investigations because an arbitrary minimum number of collaborators cannot be found. I firmly believe that any data properly taken are better than no data.

To properly interpret the analytical results of a collaborative test, the scientist must give careful consideration to the various sources of error in the data obtained. Any analytical result is a complex of three factors: (1) the random error; (2) the inherent systematic error in the procedure; and (3) the modification in this systematic error that is a consequence of any particular laboratory's environment, equipment, and any personal way of using the procedure.

One modern trend is the increasing use of reference materials and the adjustment of instruments to make them deliver the known values for the reference materials. Properly used, the reference materials delete the second and third factors mentioned in the preceding paragraph. There is a disadvantage, however, because the process of adjusting the instrument also involves a random error. The standardization of a volumetric reagent is a similar situation. There is still the

random titration error in determining the titer. But this random error, which now becomes a part of the titer value, acts as a constant, or systematic error, when the reagent is used on a series of unknowns: Any error in the value of the titer is directly carried over into every result. Consequently at least three or four titrations should be made when standardizing a reagent. The average of these repeats allows a certain degree of cancelling out of the random errors. The random error in the average of four is just half that of a single titration. This reduces the random error in the titer value to the point at which, now acting as a systematic error, it is smaller than the random error of a single titration on an unknown. This averaging device is not possible in the adjustment of instruments.

Most introductory statistical texts are necessarily limited to the consideration of very simple experimental situations. After all, the statistical student has to begin his learning with simple techniques. The real world of measurement is usually an intricate place and requires careful examination by the scientist to enumerate the various sources of error in his measurement. Since a statistician without much experience may easily overlook the hidden complexities in an unfamiliar field of measurement, the scientist must not shirk the responsibility for the interpretation of his data. This is the best reason for writing this manual, which represents the initial action in filling a recognized need for statistical guidelines in analytical method studies. Comments and suggestions from the users are invited. Further developments along these lines are intended, and revisions, additions, and deletions will be made as experience dictates.

Published by the Association of Official Analytical Chemists, Inc., Washington, D.C., 1967.

Subject Index[1]

[Reference is to volume number, boldface, followed by page number of this volume.]

[1]The letter "A" following the page number indicates an abstract only.

D

E

M

N

O

P

Q

R

S

T

Author Index

☆ U.S. GOVERNMENT PRINTING OFFICE : 1979 O—290-039

Announcement of New Volumes in the NBS Special Publication 300 Series Precision Measurement and Calibration

Superintendent of Documents,
Government Printing Office,
Washington, D.C. 20402

Dear Sir:

Please add my name to the announcement list of new volumes to be issued in the series: National Bureau of Standards Special Publication 300, Precision Measurement and Calibration.

Name _____

Company _____

Address _____

City _____ State _____ Zip Code _____

(Notification key N-353)

(cut here)

Official SI Unit Names and Symbols

[For a complete statement of NBS practice, see NBS Tech. News Bull. Vol. 52, No. 6, June 1968.]

Name	Symbol	Name	Symbol
meter	m	newton	N
kilogram	kg	joule	J
second	s	watt	W
ampere	A	coulomb	C
kelvin[1]	K	volt	V
candela	cd	ohm	Ω
radian	rad	farad	F
steradian	sr	weber	Wb
hertz	Hz	henry	H
lumen	lm	tesla	T
lux	lx		

Additional Names and Symbols approved for NBS use

curie[2]	Ci	mho	mho
degree Celsius[3]	°C	mole	mol
gram	g	siemens[4]	S

[1] The same name and symbol are used for thermodynamic temperature interval. (Adopted by the 13th General Conference on Weights & Measures, 1967.)

[2] Accepted by the General Conference on Weights & Measures for use with the SI.

[3] For expressing "Celsius temperature"; may also be used for a temperature interval.

[4] Adopted by IEC and ISO.

Table for Converting U.S. Customary Units to Those of the International System (SI)[5]

To relate various units customarily used in the United States to those of the International System, the National Bureau of Standards uses the conversion factors listed in the "ASTM Metric Practice Guide", NBS Handbook 102. These are based on international agreements effective July 1, 1959, between the national standards laboratories of Australia, Canada, New Zealand, South Africa, the United Kingdom, and the United States.

To convert from:

(1) inches to meters, multiply by 0.0254 exactly.
(2) feet to meters, multiply by 0.3048 exactly.
(3) feet (U.S. survey) to meters, multiply by 1200/3937 exactly.
(4) yards to meters, multiply by 0.9144 exactly.
(5) miles (U.S. statute) to meters, multiply by 1609.344 exactly.
(6) miles (international nautical) to meters, multiply by 1852 exactly.
(7) grains (1/7000 lbm avoirdupois) to grams multiply by 0.064 798 91 exactly.
(8) troy or apothecary ounces mass to grams, multiply by 31.103 48...
(9) pounds-force (lbf avoirdupois) to newtons, multiply by 4.448 222...
(10) pounds-mass (lbm avoirdupois) to kilograms, multiply by 0.453 592...
(11) fluid ounces (U.S.) to cubic centimeters, multiply by 29.57...
(12) gallons (U.S. liquid) to cubic meters, multiply by 0.003 785...
(13) torr (mm Hg at 0 °C) to newtons per square meter, multiply by 133.322 exactly.
(14) millibars to newtons per square meter, multiply by 100 exactly.
(15) psi to newtons per square meter, multiply by 6894.757...
(16) poise to newton-seconds per square meter, multiply by 0.1 exactly.
(17) stokes to square meters per second, multiply by 0.0001 exactly.
(18) degrees Fahrenheit to kelvins, use the relation $t_K = (t_F + 459.67)/1.8$.
(19) degrees Fahrenheit to degrees Celsius, use the relation $t_C = (t_F - 32)/1.8$.
(20) curies to disintegrations per second, multiply by 3.7×10^{10} exactly.
(21) roentgens to coulombs per kilogram, multiply by $2.579\,760 \times 10^{-4}$ exactly.

[5] Système International d'Unités (designated SI in all languages).